Access to and Retrievability of Content in Web Archives

Thaer Mahmoud Hasan Samar

Access to and Retrievability of Content in Web Archives

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 30 oktober 2017
om 10.30 uur precies

door

Thaer Mahmoud Hasan Samar

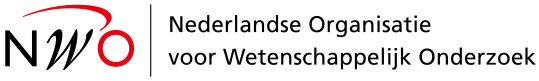
geboren te Ramallah, Palestina

Promotor:	Prof. dr. ir. A. P. de Vries	
Manuscriptcommissie:	Prof. dr. ir. Th.P. van der Weide	
	Prof. dr. W. Nejdl	(Leibniz Universität Hannover, Duitsland)
	Prof. dr. F.M.G. de Jong	(Universiteit Utrecht)
	Prof. dr. J.J. Noordegraaf	(Universiteit van Amsterdam)
	Dr. ir. J. Kamps	(Universiteit van Amsterdam)

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research reported in this thesis has been mostly carried out at CWI, the Dutch National Research Laboratory for Mathematics and Computer Science, within the Information Access group.

SIKS Dissertation Series No. 2017-32
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



This work is part of the research programme Continuous Access To Cultural Heritage (CATCH). With project number #640.005.001, which is financed by the Netherlands Organisation for Scientific Research (NWO).

Printed by Ipskamp Printing.

ISBN: 978–90–9030500–4

To my family.

ACKNOWLEDGEMENTS

Now, at the end of my PhD, I would like to thank everyone who made me enjoy doing my PhD and supported me during my PhD journey.

- First of all, I would like to thank my promotor Arjen. You were always around when I needed you, you gave me the support and the courage to find my path in research. Many thanks for supervising my PhD, which including writing research papers, formulating research question, how to present the work, and the support to finish writing the thesis.
- I would like to thank all members of the WebART project. Many thanks for the project leader Jaap Kamps. I would like to thank Hugo who doing a PhD as well, thanks Hugo for editing the Dutch summary of my thesis.
- I would like to thank all my co-authors who collaborated with me in different projects during my PhD, many thanks for your valuable contribution.
- I would like to thank my colleagues in the Information Access group at CWI.
- I would like to thanks Jacco and Lynda for reviewing my Introduction and Conclusion chapters, they made sure not to give me comments on paper but to discuss every single point.
- I would like to thank NWO-CATCH for supporting me during me research.
- Most of the work in this thesis required a high computation power, I would like to thank SURFSara and Vancis for making their resources available for us.
- Many thanks go to my family, my parents, my brothers, my sisters and my family in law. Their regular calls and check were always helpful and motivating. I am very grateful to my parents for their support and courage during the entire journey of my education, from school to PhD.
- I would like to thank the person who accompanied me in my PhD journey and decided to be abroad with me, my wife. Thanks for supporting and encouraging me. I also thank my kids Mahmoud (5 years), Mariam (2 year), and Yamin (5 months), you always made me happy and proud father.

CONTENTS

1 INTRODUCTION 1
- 1.1 Research Questions 2
- 1.2 Thesis Structure 5
- 1.3 Publications 6

2 RELATED WORK 9
- 2.1 Web Archiving 10
  - 2.1.1 Web Archiving Projects 13
- 2.2 Web Archives Completeness 16
- 2.3 Link Structure and Anchor Text 17

I ACCESSIBILITY OF WEB ARCHIVE CONTENT ALONG THE TIME AXIS – STUDYING A LARGE-SCALE WEB ARCHIVE COLLECTION 23

3 UNCOVERING THE UNARCHIVED WEB 25
- 3.1 Introduction 25
- 3.2 Experimental Setup 27
  - 3.2.1 Data 27
  - 3.2.2 Link Extraction 28
  - 3.2.3 Link Aggregation 29
- 3.3 Expanding the Web Archive 29
  - 3.3.1 Archived Content 29
  - 3.3.2 Unarchived Content 31
  - 3.3.3 Characterizing the "Aura" 33
- 3.4 Representations of Unarchived Content 34
  - 3.4.1 Indegree 35
  - 3.4.2 Anchor Text Representations 36
  - 3.4.3 Homepage Representations 36
  - 3.4.4 Qualitative Analysis 37
- 3.5 Finding Unarchived Pages 38
  - 3.5.1 Evaluation Setup 39
  - 3.5.2 Availability of Pages 40
  - 3.5.3 MRR and Success Rate 40
  - 3.5.4 Impact of Indegree 42
- 3.6 Discussion and Conclusions 43

4 TEMPORAL ANCHOR TEXT AS PROXY FOR PAST USER QUERIES 45
- 4.1 Introduction 45
- 4.2 Setup 46
  - 4.2.1 Dataset 46
  - 4.2.2 Link Extraction & Aggregation 46
  - 4.2.3 Wikipedia Page Views Statistics 47
- 4.3 Analysis 48

4.3.1 Hosts Evolution 48
4.3.2 Anchor Text Evolution 48
4.3.3 Matching Anchor Text To Wikipedia Title 49
4.4 Conclusions and Future Work 54
5 COMPARING TOPIC COVERAGE IN BREADTH-FIRST & DEPTH-FIRST CRAWLS 57
5.1 Introduction 57
5.2 Setup 59
5.2.1 Data 59
5.2.2 Anchor Links Extraction 59
5.2.3 Link Subsets from *Common Crawl* 60
5.2.4 Sources of Topics 61
5.3 Analysis 62
5.3.1 Target Pages 63
5.3.2 Anchor Text 65
5.3.3 Topic Coverage 65
5.4 Conclusions 70
6 QUANTIFYING RETRIEVAL BIAS IN WEB ARCHIVE SEARCH 73
6.1 Introduction 74
6.2 Related Work 76
6.3 Approach 78
6.4 Experimental Setup 79
6.4.1 Retrievability Experimental Setup 79
6.4.2 Known-Item Search Setup Based on Retrievability Scores 85
6.5 Retrievability Bias 88
6.5.1 Retrievability and Findability 91
6.6 Impact of Number of Versions on the Retrievability Bias 92
6.6.1 Collapsing Similar Versions 92
6.6.2 Collapsing Versions (URL-based) 96
6.7 Quantification of Retrieval Bias Over the Years 97
6.7.1 Time-based Subsets based on Time-based Queries 99
6.8 Discussion & Conclusions 101

II OPEN WEB (LIVE AND DYNAMIC) & CRAWLED WEB (ARCHIVED AND STATIC) 105

7 THE STRANGE CASE OF REPRODUCIBILITY VS. REPRESENTATIVENESS IN CONTEXTUAL SUGGESTION TEST COLLECTIONS 107
7.1 Introduction 108
7.2 Related Work 110
7.3 Experimental Setup 113
7.3.1 DataSet 113

7.3.2 URL Normalization 114
7.3.3 Mapping Open Web qrels to *ClueWeb12* 114
7.3.4 Expanding ClueWeb12 qrels 115
7.3.5 Mapping Open Web URLs to the *ClueWeb12* documents Ids 115
7.4 Comparing Open Web and Closed Web Relevance Assessments 115
7.4.1 Fair Comparison 116
7.4.2 Comapring Identical Documents from Open Web & *ClueWeb12* 116
7.5 Reproducibility of *Open Web* Systems 120
7.6 Conclusions 124
8 IMPROVING CONTEXTUAL SUGGESTIONS USING OPEN WEB DOMAIN KNOWLEDGE 125
8.1 Introduction 125
8.2 Experimental Setup 127
8.3 Contextual Suggestion Model 127
8.3.1 General Model and Problem Formulation 127
8.3.2 Personalization 128
8.3.3 Selection Methods of Candidates 128
8.4 Sub-collections Discussion 134
8.5 Effect of Using External Domain Knowledge for Candidate Selection 136
8.6 Insights on the Results 137
8.6.1 Analysis per RelevanceDimensions 137
8.6.2 Impact of used Filters 139
8.6.3 Effect of Prior Probability 139
8.7 Conclusion 141
9 CONCLUSIONS 143
9.1 Main Findings 144
9.2 Future Work 154

BIBLIOGRAPHY 159
Summary 175
Samenvatting 179
List of Publications 181
List of Figures 185
List of Tables 188
SIKS Dissertation Series 195

1 INTRODUCTION

The World Wide Web (*WWW* or simply the *Web*) offers a rich means for its users to publish, share, create, discuss, collaborate and even earn a living. The diversity and magnitude of data created makes it an interesting basis for many studies on user behavior in different tasks, such as information seeking. Web content, however, is surprisingly volatile, where 80% of Web pages may disappear within one year [136]. Culture heritage institutions increasingly recognize that such digital born data are as easily deleted as they are published, thereby introducing risks to the world's digital cultural heritage [153]. In order to preserve the content on the web, many national libraries and archives have set up Web archiving initiatives. Web archives address this problem by systematically preserving parts of the Web for future generations. Web archiving involves a "process of collecting portions of the web, preserving the collections in an archival format, and then serving the archives for access and use" [107]. A Web archiving process, includes the following main tasks; selection, harvesting, storage (preservation), and access [130, 100]. Since 1996, web archiving has been performed by international and national heritage institutions, pioneered by the Internet Archive [14]. In July 2003, twelve institutions met at the National Library of France (*BnF*) to create the International Internet Preservation Consortium (IIPC) [12]. The aim of IIPC is to develop tools for archiving the web, promoting the access and use of web archives for research and cultural heritage.

Archiving the entire Web is an impossible task due its increasing size, and the dynamic and ephemeral nature of its content. However, with the effort of all Web archiving initiatives parts of the web are being preserved. Web archives are created by crawling Web pages following a crawling strategy determined by the institutions, leading to differences in scope and coverage of the Web archives. The National Library of The Netherlands (KB) [16], for example, archives websites selected on the basis of categories related to Dutch historical, social and cultural heritage. The crawling frequency varies among the selected websites: some are crawled yearly, biannually, quarterly, others on daily basis, such as a news agency.

The research presented in this thesis was conducted within the *WebART*[1] (***Web** **A**rchive **R**etrieval **T**ools*) project. WebART is part of the Dutch Continuous Access To Cultural Heritage (CATCH) research programme, funded by the Netherlands Organization for Scientific

1 http://www.webarchiving.nl/

Research (*NWO*[2]). WebART is an interdisciplinary research project, which combines expertise in Computer Science, Information Science and New Media research from the University of Amsterdam, and the Dutch national center for computer science and mathematics *CWI* (Centrum for Wiskunde and Informatica). The cultural heritage partner is the National Library of the Netherlands (KB). The WebART project aims to critically assess the value of Web archives for realistic research scenarios, and develop information access tools and methods that maximize the archive's utility for research.

1.1 RESEARCH QUESTIONS

This thesis is divided into two parts, in each part we address a number of related research questions. In the first part (*Part I – Studying Large-Scale Web Archives*), we apply a large-scale analysis on an archive collection investigating the archived and *unarchived* content. In the second part (*Part II – Integrating Online & Crawled Web*), we investigate how to link information from the current Web (*online*) to the past Web (*archived*).

Part I – Studying Large-Scale Web Archives focuses on the accessibility and retrievability of Web archive collections. We do not only focus on accessing Web pages that exist in the archive, but we also study the uncrawled Web pages; pages that existed on the Web at crawling time but have not been crawled. Web archives are known to be incomplete; it is impossible to crawl the entire Web due its increasing size and transient nature of its content, and new Web pages are added constantly, new Web page content replaces old content without preserving the old content. Therefore, we study the possibility of finding traces of *unarchived* Web pages from the archived pages, and whether they have been really lost forever because they were not archived. More precisely, we go back in time and try to *uncovering* and *reconstruct* the *unarchived* Web pages. We ask the following main research question:

RQ1 *Can we uncover and provide representations of unarchived Web pages exploiting references to them from the archived Web pages?*

In the field of Information Retrieval, *Anchor Text* has been used to enrich the representation of Web page content to improve Web search effectiveness [73, 85, 90, 110, 117, 119, 121, 132]. In the Web archive context, hyperlink structure has been used to estimate what is missing in the Web archive. Links and *Anchor Text* can be used to locate missing Web pages, of which the original URL is not accessible anymore [115]. We exploit the hyperlink structure of the crawled content to derive evidence of the existence of *unarchived* pages. When

2 http://www.nwo.nl/

we encounter references to Web pages that have not been crawled from archived pages (we call this *link evidence*). Existence of a link implies that the page existed on the Web at or before the crawling time. It is not possible to go back in time and crawl the missing pages. Instead, we show that it is possible to reconstruct unarchived pages by providing basic representation about them, instead of losing them. We build implicit representations of *unarchived* Web pages and domains, based on link evidence and *Anchor Text*, and investigate the richness (or sparseness) of the descriptions in the number of incoming links and the aggregated *Anchor Text*. In order to study whether the resulting derived representations of *unarchived* pages are useful in practice, we setup a known-item search experiment.

Chapter 3 answers RQ1.

While we can archive Web pages, queries from the past have usually not been preserved. We cannot go back in time and find out topics (queries) that were of interest to users in the past; at crawling time. *Anchor Text* exhibits characteristics similar to real user queries, and similar to titles of Web pages [86]. This is based on the observation that titles can be used as an approximation of queries [108]. We explore the use of time-aware *Anchor Text*, extracted from link structure of a Web archive collection, in order to investigate what was popular on the Web. We ask the following research question:

RQ2 *Can we identify past popular topics using anchor text associated with hyperlinks of the Web archive?*

Chapter 4 answers RQ2.

As pointed out above, it is practically infeasible to archive the entire Web due its increasing size, and the ephemeral nature of its content. Therefore, institutes have to make decisions on the websites to be included in the crawling process, the crawling frequency, and the crawling strategy. The crawling strategy followed to create a Web archive has a great influence on the data to be archived. One strategy is to crawl a manually selected set of websites (called the crawler's *seeds*) and to harvest these websites in depth (*depth-first* crawl). Another strategy automatically crawls as many websites as possible, but not in depth (*breadth-first* crawl). Both crawling strategies result in incomplete crawls, as both strategies exclude websites. *Depth-first* ignores websites outside the *seed* list, and *breadth-first* archives websites incompletely as crawling cannot follow all links to sub-pages. We investigate the influence of these crawling strategies on the coverage of topics from the past using *Anchor Text*. We address this by studying two Web crawls created with different crawling strategies. That is our third research question:

RQ3 *How does the crawling strategy impact the Web archive's coverage of past popular topics?*

Chapter 5 answers RQ3.

Indexing and retrieving documents from a Web archive collection can be difficult because multiple versions of the same document may appear in the ranked search results. Dealing with exact- or near-duplicates has been addressed in several studies, e. g. [94, 127]. Measuring retrieval effectiveness of a search system is usually done experimentally by applying evaluation measures and test collections. To complement the standard IR evaluations which focus on the assessment of efficiency and effectiveness of IR systems, Azzopardi et al. introduced a *retrievability* [39] metric to estimate the likelihood of retrieving a document by a specific retrieval system by issuing a large set of queries and analyzing the result sets. We investigate how we can quantify retrieval bias in Web archives. Specifically, we ask the following main research question:

RQ4 *What can we learn about Web archive access from studying the collection using a measure of retrievability?*

Chapter 6 answers RQ4.

In Part II – Integrating Online & Crawled Web of the thesis, we investigate how to link information from the current Web (*online*) to the past Web (*archived*). Our approach to investigate this research question resulted in participating in the TREC Contextual Suggestion (*CS*) track. The *CS* track provides an evaluation framework for systems that recommend items to users given their geographical context. The specific nature of this track allows the participating teams to identify candidate documents either from the *Open Web* or from the *ClueWeb12* collection, a static crawl from the Web. In the 2013 and 2014 editions of the *CS* track, submissions based on the *Open Web* outnumbered those based on the *ClueWeb12* collection. However, to achieve reproducibility, ranking web pages from *ClueWeb12* should be the preferred method for scientific evaluation of contextual suggestion systems. It has been found that the systems that build their suggestion algorithms on top of input taken from the *Open Web* achieve consistently a higher effectiveness than systems based on the *ClueWeb12* collection. Most of the existing works have relied on public tourist APIs to address the contextual suggestion problem. These tourist sites (such as Yelp and Foursquare) are specialized in providing tourist suggestions, hence those works are focused on re-ranking the resulting candidate suggestions based on user preferences.

The finding that Open Web results achieve higher effectiveness raises the question whether research systems built on top of the *ClueWeb12* collection are still representative of those that would work directly on industry-strength web search engines. Therefore, we focus on analyzing reproducibility and representativeness of *Open Web* and *ClueWeb12* systems.

We study the gap in effectiveness between *Open Web* systems and *ClueWeb12* systems through analyzing the relevance assessments of documents in each set, and overlap. In the judging pool of judged documents, the documents from the *Open Web* and *ClueWeb12* collection are distinguished. Hence, each system submission should be based only on one resource, either *Open Web* (identified by URLs) or *ClueWeb12* (identified by ids). We ask the following question:

RQ5 *Do relevance assessments of Open Web differ (significantly) from relevance assessments of ClueWeb12 documents? Can we identify an overlap between the two sets, and the documents in the overlap were judged?*

Chapter 7 answers RQ5.

We propose an approach for selecting documents from *ClueWeb12* collection based on information obtained from location-based social networks on the *Open Web*. This makes an improvement step towards partially bridging the gap in effectiveness between *Open Web* and *ClueWeb12* systems, while at the same time we achieve reproducible results on well-known representative sample of the web. We mainly ask the following research questions:

RQ6 *Can we identify a representative sample from the ClueWeb12 collection by applying filters from the Open Web tourist APIs tailored for the CS track?*

Chapter 8 answers RQ6.

1.2 THESIS STRUCTURE

This thesis consists of six research chapters organized in two parts. Part I – Studying Large-Scale Web Archives consists of Chapters 3, 4, 5, and 6, while Part II – Integrating Online & Crawled Web consists of Chapters 7, and 8. Chapter 2 provides the related work of research chapters. Each chapter anwers one of the main research questions described in the previous section. The two parts can be read independently. Chapter 9 draws the thesis conclusions, the graphical representation of the thesis structure is shown in Figure 1.

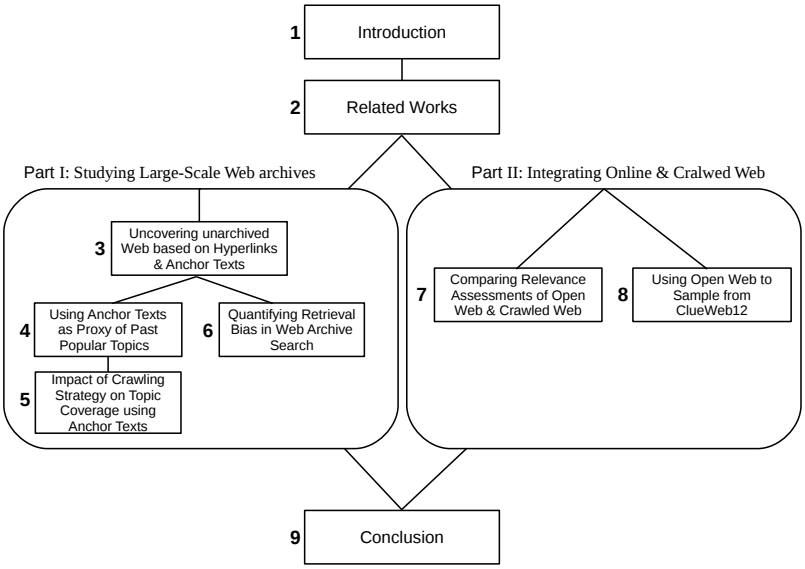


Figure 1: **Thesis Structure.**

1.3 PUBLICATIONS

This section presents the list of publications that contribute to the thesis.

Part I – Studying Large-Scale Web Archives- Accessibility of Web Archive Content Along the Time Axis

Chapter 3

Thaer Samar and Hugo C. Huurdeman and Anat Ben-David and Jaap Kamps and Arjen P. de Vries. *Uncovering the* unarchived *web*. The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, Gold Coast, QLD, Australia, 2014.

Hugo C. Huurdeman and Anat Ben-David and Jaap Kamps and **Thaer Samar** and Arjen P. de Vries. *Finding pages on the* unarchived *Web*. IEEE/ACM Joint Conference on Digital Libraries, JCDL, London, United Kingdom, 2014.

Hugo C. Huurdeman and Jaap Kamps and **Thaer Samar** and Arjen P. de Vries and Anat Ben-David and Richard A. Rogers. *Lost but not forgotten: finding pages on the* unarchived *web*. International Journal on Digital Libraries, 2015.

Chapter 4

Thaer Samar and Arjen P. de Vries. *Temporal Anchor Text as Proxy for Real User Queries*. Proceedings of the 5th International Workshop on Semantic Digital Archives SDA, co-located with the International Conference on Theory and Practice of Digital Libraries, TPDL, Poznań, Poland, 2015.

Chapter 5

Thaer Samar and Myriam C. Traub and Jacco van Ossenbruggen and Arjen P. de Vries. *Comparing Topic Coverage in Breadth-first & Depth-first Crawls using Anchor Texts*. Research and Advanced Technology for Digital Libraries - 20th International Conference on Theory and Practice of Digital Libraries, TPDL, Hannover, Germany, 2016.

Chapter 6

Thaer Samar and Myriam C. Traub and Jacco van Ossenbruggen and Lynda Hardman and Arjen P. de Vries. *Quantifying Retrieval Bias in Web Archive Search*. *International Journal on Digital Libraries, 2017.* .

Part II – Integrating Online & Crawled Web- Open Web (live and dynamic) & Crawled Web (archived and static)

Chapter 7

Thaer Samar and Alejandro Bellogín and Arjen P. de Vries. *The Strange Case of Reproducibility vs. Representativeness in Contextual Suggestion Test Collections*. Information Retrieval Journal, 2015.

Alejandro Bellogín and **Thaer Samar** and Arjen P. de Vries and Alan Said. *Challenges on Combining Open Web and Dataset Evaluation Results: The Case of the Contextual Suggestion Track*. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR, Amsterdam, The Netherlands, 2014.

Chapter 8

Thaer Samar and Alejandro Bellogín and Arjen P. de Vries. *Improving Contextual Suggestions using Open Web Domain Knowledge*. Proceedings of the International Workshop on Social Personalisation & Search, SPS, co-located with the 38th Annual ACM SIGIR Conference (SIGIR 2015), Santiago de Chile, Chile, August , 2015.

The publications supporting Chapter 7 and Chapter 8 are based on the two participation in the TREC Contextual Suggestion (CS) track; CS 2013 and CS 2014. The technical publication appeared in the proceeding of the Text REtrieval Conference TREC.

CS 2013

CWI and TU Delft *Notebook TREC 2013: Contextual Suggestion*, Federated Web Search, KBA, and Web Tracks. Alejandro Bellogín and Gebrekirstos G. Gebremeskel and Jiyin He andAlan Said and **Thaer Samar** and Arjen P. de Vries and Jimmy Lin and Jeroen B. P. Vuurens. Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, 2013.

CS 2014

Better Contextual Suggestions in ClueWeb12 Using Domain Knowledge Inferred from The Open Web. **Thaer Samar** and Arjen P. de Vries and Alejandro Bellogín. Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, 2014.

2

RELATED WORK

The content on the Web is very dynamic, pages are added, deleted, and modified continuously. In 2004, Ntoulas et al. study the dynamic nature of a Web crawl, they showed that 80% of Web pages are not accessible after one year [136]. In the literature, there are several studies that investigated the Web dynamics. The experimental results on the evolution of Web pages over time were used to propose an architecture for the incremental crawler [65]. In order to study the evolution of the Web, they employed active crawling of 720,000 Web pages on daily basis over four months. They found that 15% of the Web pages have a change interval longer than one day and shorter than a week, 20% of Web pages changed daily, and the rate of change depends on the pages top-level domains. For example, more than 40% of pages that change daily are from *.com* top-level domain. Web pages from *.edu and .gov* top-level domains are very static, more than 50% of pages from these domains did not change at all in four months of crawling. This study was expanded by [88] in terms coverage and the rate and degree of change. They crawled 150,836,209 Web pages on a weekly basis over a span of 11 weeks. They found a strong relationship between the top-level domains and the frequency of change of a Web page, and a weak relationship between the top-level domains and the degree of change. The largest study of Web evolution was carried by [104] over an 18 years period (from 1996 to 2013) of the German Web part archived by the *Internet Archive*. Based on the evolution analysis of popular domains (websites and their sub-domains) from current Web Amazon's Alexa ranking [2], they observed that 70% of Web pages are younger than a year, and popular domains have growing exponentially. Many institutions realized the importance of collecting and preserving the Web pages before they are lost. Many projects have been initiated to preserve the Web, build tools for facilitating access to Web archives. In the following sections, we summarize the Web archiving process and tasks involved in this process, introduce some of the projects that are involved in crawling the Web and making the crawls available for researchers as well as to the public, or projects building tools to facilitate access to Web archives.

In this chapter, we discuss the background for the research presented in this thesis. First, we introduce the Web archiving process and present projects involved in archiving the Web (Section 2.1). Then, we discuss background of research related to the completeness of Web archives (Section 2.2). The hyperlinks and Anchor Text play an important role in the thesis. In Section 2.3, we introduce hyperlinks

and Anchor Text, and work related to them in the context of Web retrieval and Web archiving.

2.1 WEB ARCHIVING

Web archives process consists of the following tasks: selecting, harvesting, storing, preserving, and providing access to archived Web content over time [130, 100].

Selection

The selection task is the process of selecting and deciding which websites (*seeds*) to be collected from the Web using Web *crawlers*. Web archives use different sources to maintains the seeds list based on their goals. Many national libraries focus on archiving their national Web domain fully or partially, this type of archiving is called domain archiving. For example, the National Library of the Netherlands (KB) [16] archives a manually selected list of websites [140] of value for Dutch heritage. Web sites for preservation are selected by the library per categories related to Dutch historical, social and cultural heritage. Each selected website has been assigned a UNESCO code corresponding to the category to which it belongs. The National Library of France (*BnF*) [6] performs a broad crawl of the entire french domain (*.nl*) in addition to the selection-based crawl. Internet Archive [14] collects websites from the whole Web that are publicly available and not excluded by *robot.txt* [24]. Other Web archives are focused around a specific topic or event. For example, the UK Web archive special collections [27] created by the British Library [26]. These collections are collections of websites grouped together on a particular theme by librarians. These collections can be event-based (e. g.*UK General Election*), or topical-based (e. g.*The Credit Crunch Collection*), or subject-based (e. g.*The British Countryside Collection*).

Harvesting

The harvesting task or *crawling* is referred to the process of getting content from the Web into archives. The programs collecting the Web content are called Web *crawlers*. A crawler starts with the given list of websites *seeds* as starting point. First, it retrieves the content of the provided *seeds*. Then, extracts any hyperlink and adds the URL to the queue of URLs to be harvested. The *Web crawling* has been widely studied in the literature. *Web crawlers* are an important component of Web search engines. Web search engines aim to maintain an index of the Web and to give access to a recent copy of Web pages. In the context of Web archiving, *Web crawlers* play an important role as they are used by institutions to periodically crawl the Web and store it

in the archives. *Web crawling* research involves different issues and challenges, such as managing large dataset, the increasing size and the dynamic nature of the Web which has an impact on the coverage and freshness of the crawled content. These challenges are others are discussed in [137], which is a survey about the crawling algorithms and strategies, presents a chronology of *Web crawlers* development, and outlines the fundamental challenges.

Storage

Web crawlers register additional information (*metadata*) about the harvested Web page such as the crawling timestamp. Collected material are stored on a storage medium.

Preservation

Preservation is the process of saving the digital content of archived Web pages and ensure a continued accessibility over time. Therefore, to achieve this goal tools, standards are needed, for example, standards for archival format. The *ARC* [60] file format has been developed in 1996 by Brewster Kahle and Mike Burner from the Internet Archive for storing and managing a large number of objects harvested from Web as sequence of content blocks. Each block consists of a metadata header which contains information about the crawl such as the URI and the timestamp of the crawling date, and of the raw content of the crawled object. The *WARC* (*Web ARChive*) file format [29] is an extension of the *ARC* format.

Access

In early stages of the Web archiving initiatives, the main focus was on developing tools for collecting and preserving the Web content, with less attention to the use of the Web archives [130, 102]. However, in addition to preserving Web data for the future, Web archives provide a rich data source for researchers as well as for the general public. Web archives should be available for access and use as stated by the International Internet Preservation Consortium (IIPC). Web archives are increasing getting attention of researchers from different disciplines. For example, a survey was conducted by the *UK* Web archive to collect information about the scholarly use of their archive. The majority of these researchers (94) are from Arts, Humanities, and Social Sciences disciplines [102]. In the same study, they observed a significant increase on the usage (such as number of users and page views) of the *UK* Web archive in 2013 compared to the usage in 2012. The increase was noticed since April 2013, when the non-print legal

deposit regulations were effective in the *UK* and the Web archive was frequently mentioned in the media.

In 2001, the Internet Archive[1] made their Web archive accessible to the public through the Wayback Machine. However, accessing Web archives through the Wayback Machine is limited as it requires the user to provide the URL for search and then view its archived snapshots over time. The open-source Wayback Machine[2] has been widely used by the Web archive initiatives to provide access to their collections.

Since the release of Internet Archive's Wayback Machine in 2001, searching for the URL has been the main way for accessing Web archives. Recently, Web archive initiatives started to provide *full-text* search. In a survey conducted by the National Library of the Netherlands (KB), they reported that *full-text* search functionality was ranked first in the list of top ten functionalities that users of the KB Web archive would like to be implemented [141]. In order to understand the information needs of Web archive users, data was collected from users of the Portuguese Web Archive (*PWA*) [95], which is publicly available since 2010 through search interface that allow users to perform *full-text* and *URL* search [23]. Three methods were used to collect the users data: search logs, an online questionnaire, and a laboratory study [67]. The results of these methods were coincident, they found that users prefer *full-text* search over *URL* search. Users perform navigational search without time restrictions, this result was inline with the KB survey, the time functionality wan not among the top ten functionalities that users would like to have. The time functionality was one of the least frequently used. However, they observed that when the time functionality was used, it was used to get the oldest documents.

The shift from single URL search to search interfaces was described as a turning point in the history of Web archives [50]. Web archive initiatives started to allow access to their archives through *full-text* search using existing IR retrieval systems. Through a survey conducted in 2010 of 42 Web archives initiatives across 26 countries [96]. They found that 89% of the initiatives support access to the Web archive of a given URL. For instance, Internet Archive *WayBack Machine* [30] allows users to navigate and browse captures versions of Web pages over time, the user has to know the *URL* in advance. 79% enable searching meta-data and 67% provide *full-text* search for the entire or part of their archived collection. The same survey was conducted again in 2014 in order to observe the change in Web archiving since 2010 [72]. In terms of access methods, the results of 2014 are the same as in 2010.

1 https://archive.org/
2 http://archive-access.sourceforge.net/projects/wayback/

Figure 2: The Internet Archive Wayback Machine URL-based search interface. The screen shot was taken on September 1 2016 at 11:55 AM (CEST).

2.1.1 *Web Archiving Projects*

In this section, we present projects initiated for improving Web archive research or are active in crawling the Web and making collections of snapshots from the Web available for research.

The **Internet Archive** is a non-profit foundation. In 1996, the Internet Archive took the initiative to archive the entire Web with the goal to build an Internet library and to make it accessible to the public. The Internet Archive's Wayback Machine [30] is the oldest and largest Web archive that is available for public search, the unit of search is a URL (see Figure 2). The user enters a URL of interest and gets a calendar showing the crawls of the given URL over time (see Figure 3). When the user clicks on a specific timestamps, the content of the URL will be shown at was on that time. Internet Archive collaborates with various national libraries to help them archives there national domains. In 2006, Internet Archive created *Archive-It* [4] which a Web archiving subscription service that helps organizations to crawl, build and preserve collections of digital content. These organization are able to manage the content of their archives (hosted on the Internet Archive data centers) with full accessibility using a Web-based application, a full-text search is available for them and their patrons.

Followed by the Internet Archive, different parts of the Web has been preserved by different initiatives world wide [21], such as the national libraries. Many of these institutes are members in the International Internet Preservation Consortium (IIPC) [12]. IIPC was established in 2003 with the goal to create a collaboration between organizations who are doing Web archiving world wide including museums, libraries, national libraries, and culture heritage institutions. The mission of the IIPC is to acquire, preserve, and provide accessibility of the preserved data for future generation [12]. Many tools have been developed to support different stages of the Web archiving process such as tools for crawling, for example *Heritrix* [11] software, tools for accessing Web archives, indexing and searching tools such as *NutchWAX (Nutch with Web Archive eXtensions)* and *WERA (WEb aRchive Access)*, a complete list of tools that are recommended and used by members of the IIPC [13].

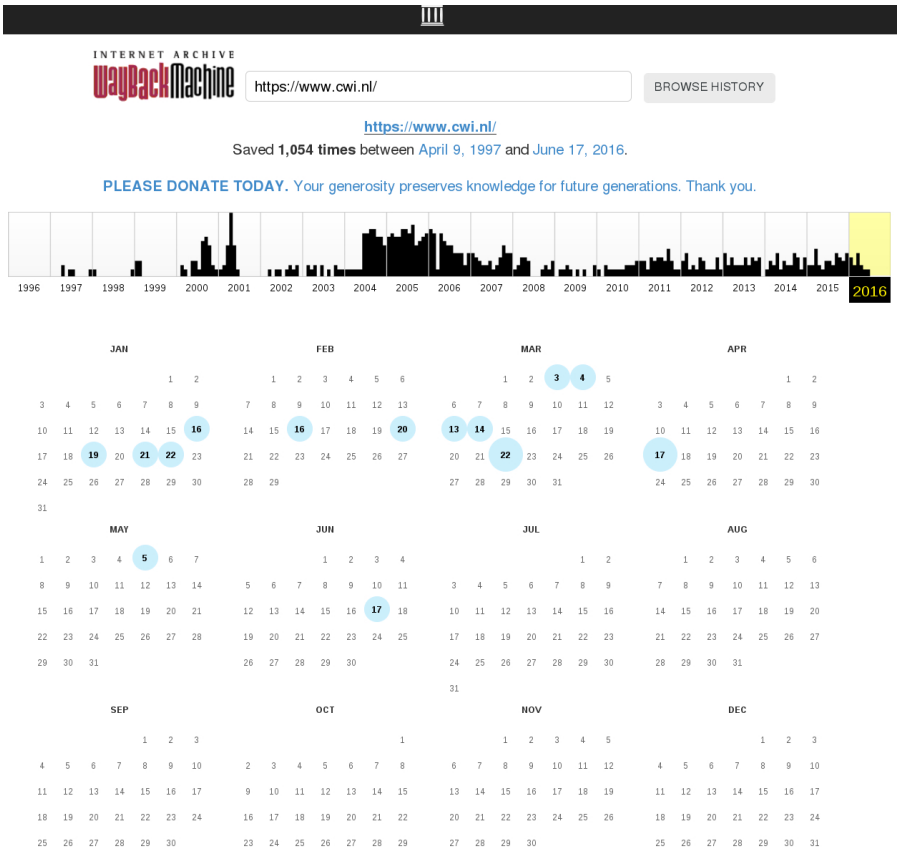


Figure 3: Crawls of `https://www.cwi.nl/` returned by the Internet Archive Wayback Machine.

There are some projects which collect data from the Web for the purpose of making data available to researchers as well as to the public such as *Common Crawl* [7]. *Common Crawl* is a non-profit organization that crawls data from the Web and makes it available for everyone. The availability of Web archives data lead to initiation of research projects. The *Living Web Archives (LiWA)* [18] project (from February 2008 until January 2011) focused on six research cases of Web archiving [55]: Archive fidelity, Web spam filtering, archive coherence, archive interpretability, streaming application, and social web application. The *LiWA* project was followed by the *Longitudinal Analytics of Web Archive data (LAWA)* [17] project (from September 2010 until August 2013) aimed to develop infra-structure and tools for aggregating, querying and analyzing large-scale Web archive data. The *Archiving Community Memories (ARCOMEM)* project aimed to develop innovative models and tools for selecting, preserving and exploiting the social Web. The *Memento* [155, 154, 19] project added the temporal dimension to the *HTTP* (HyperText Transfer Protocol) which allows users to go back in time and browse the past Web. Users can send requests to obtain the archived versions (called *Mementos*) of specified *HTTP* address and the desired date after adding

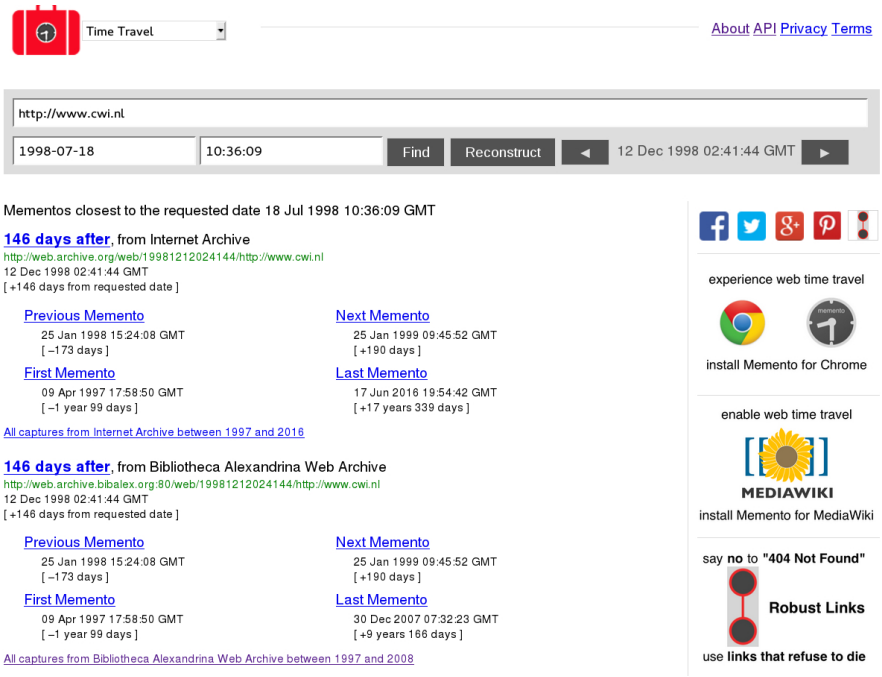


Figure 4: Mementos of `https://www.cwi.nl/` returned by the Memento Time Travel portal. One Memento per archive which holds at least one Memento.

the *Memento Time Travel* plug-in to their Web browser. Then, users will seamlessly be able to access the *Mementos* if they exist on the Web server that that holds the original archived resources and supports the *Memento* protocol. Otherwise, the *Mementos* will be served by the Web archive which has the highest coverage of the requested resource around the specified date. The *Memento Time Travel* portal [20] allow users to *find* and *reconstruct Mementos* that can be found in Web archives or in systems that support versioning such as wikis and revision control systems, which natively support the *Memento Time Travel for the Web* protocol. Using the *find* service, a list of *Mementos* for the provided *HTTP* address and the given date and time will be presented; for archives that hold at least one *Memento*, one entry will be shown per archive, entries will be sorted based on how *Mementos* are close to the specified date and time (see Figure 4).

The *ALEXANDRIA* project [3] aims to develop models and tools to explore and analyse Web archives. The main goals of the project is develop time-aware entity based enrichment and indexing, efficient indexing, retrieving and exploration of entities and events from the past. The *Warcbase* project [125, 28] is an open-source platform for storing and managing Web archives built on *Hadoop* [9] and *HBase* [10]. *Warcbase* was originally developed to take advantage of *HBase* by ingesting Web archive records into *HBase* table and allows an efficient temporal browsing using *URL* search (similar to Wayback machine). *Warcbase* provides tools for analyzing Web archives using *Hadoop/Mapre-*

duce, *Apache Pig* [22], and *Apache Spark* [25]. The *ArchiveSpark* [103, 5] project is an *Apache Spark* framework that facilitates accessing, data extraction, and analysis of Web archives.

2.2 WEB ARCHIVES COMPLETENESS

Experts in the Web archiving community discuss the shortcomings of Web archiving crawlers in terms of the content they fail to capture [129]. Some websites are intentionally excluded, breadth-first crawls might not capture deeper pages of a website, and selective crawlers exclude sites beyond the scope of the selection policy. However, as argued by Day [76], in most cases, even the sites that meet selection guidelines on other criteria may include errors, be incomplete or have broken links. Moreover, Web archiving crawlers often times fail to capture specific content elements such as JavaScript, Flash, and database-driven sites [76, 101, 129]. This prompts Web historian Brügger [56] to argue that almost every Web archive is incomplete to the extent that it is hard to determine what is missing. Brügger [57] described different levels of missing information from Web archives: Web elements level such as images, sounds, and videos that might not have been archived due to technical reasons, missing entire Web pages from Web archive. The third is missing information about the Web as a whole; information that were available when the archived content was online on the Web such as search engine results, queries, or open directories that provide statistics about the Web, for example the *Open Directory project* of the Web (*DMOZ*) [8], and *Internet World Stats* [15] which provide up to date statistics about the Web.

The limits of Web archives' crawlers may result in partial and incomplete Web archives. However, crawlers do encounter and register additional information about a page they encounter, such as its outlinks, Anchor Text, and crawl and page timestamps. Rauber et al. [142] have recognized the wealth of additional information contained in Web archives which can be used for analytical purposes. Gomes and Silva [92] used data obtained from the domain crawl of the Portuguese Web archive to develop criteria for characterizing the Portuguese Web. More recently, researchers from the LiWA project have developed a prototype for an analytical user interface designed to use these elements for analyzing large scale Web archives [151]. The Memento project has expanded the scope of analysis of archived web data beyond the boundaries of a single archive, in order to profile and analyze coverage of archived websites across different web archives. Memento [154] is an HTTP-based framework which makes it possible to locate past versions of a given Web resource through an aggregator of resources from multiple Web archives. In a recent study, Alsum et al. [36] queried the Memento aggregator to profile and evaluate the coverage of twelve public Web archives. They found that the number

of queries can be reduced by 75% by only sending queries to the top three Web archives. Here, coverage (i.e. whether a resource is archived and in which archive its past versions are located) was calculated based on the HTTP header of host level URLs.

2.3 LINK STRUCTURE AND ANCHOR TEXT

One of the defining properties of the Internet is its hyperlink-based structure. The Web's graph structure is well studied, the first classic work about the structure of the whole Web was published by Broder et al. [54] in 2000. One of the main findings was the *bow-tie* structure of the Web graph; a giant strongly connected component containing *28%* of the nodes. They showed that the in-degree distribution, the out-degree distribution, and the distribution of the sizes of the strongly connected components are heavily tailed and followed the power law. They used the AltaVista crawl of *200* million pages and *1.5* billion links. They tested and confirmed this result on a second AltaVista crawl. Serrano et al. [149] analyzed four crawls gathered between 2001 and 2004 by different crawlers with different parameters. Their main observation is that several Web crawl properties are dependent on the crawling process. Other studies of Web crawls based on different regional crawls gathered using different crawlers showed different pictures of the Web graph [84, 44, 162]. The structure of the Web was revisited and studied in [133] at scale using a Web crawl harvested by Common Crawl Foundation [7]. They found that some graph features observed by Broder et al. [54] depend on the crawling process, while other features appear to be more structural. They confirmed the existence of a giant strongly connected components, however they observed different properties of nodes that can reach or that can be reached by the giant component, suggesting that *bow-tie* structure is strongly dependent on the crawling process.

Methods to use this structure have widely been applied, especially in the context of Web retrieval (for example PageRank [122] and HITS [116]). The links which weave the structure of the Web consist of a source URL, a destination URL, and Anchor Text which is the text used to describe the target page in the link.. Aggregating Anchor Text of links makes it for example possible to create representations of target pages.

Anchor Text is a well-known resource to enrich the representations of web page content to improve Web retrieval. Eiron and McCurley [86] have investigated the properties of Anchor Text in a large intranet corpus in order to understand why using Anchor Text improves the quality of Web search. First, they showed empirically that Anchor Text exhibits characteristics similar to real user queries. Second, they hypothesize that Anchor Text is similar to web page titles, based on the observation by Jin et al. [108] that titles can be used as an approx-

imation of queries. They found that Anchor Text is indeed similar to documents titles. Craswell et al. [73] explored the effectiveness of Anchor Text in the context of site finding. Aggregated Anchor Text for a link target were used as surrogate documents, instead of the actual content of the target pages. Their experimental results show that Anchor Text can be more effective than content words for navigational queries (i.e. site finding). Work in this area led to advanced models that combine various representations of page content, Anchor Text, and link evidence [110]. Fujii [90] presented a method for classifying queries into navigational and informational. Their retrieval system used content-based or anchor-based retrieval methods, depending on the query type. Based on their experimental results, they concluded that content of webpages is useful for informational query types, while Anchor Text information and links are useful for navigational query types. Contrary to previous work, Koolen and Kamps [119] concluded that Anchor Text can also be beneficial for ad hoc informational search, and their findings show that Anchor Text can lead to significant improvements in retrieval effectiveness. They also analyze the factors influencing this effectiveness, such as link density and collection size. In the context of Web archiving, link evidence and Anchor Text could be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [115] computed lexical signatures of lost webpages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost webpages. Anchor Text can also be used for other purposes, for example for query suggestions.

Following Kleinberg [117], Dou et al. [85] took the relationships between source pages of Anchor Text into account. Their proposed models distinguish between links from the same website and links from related sites, to better estimate the importance of Anchor Text. Similarly, Metzler et al. [132] smoothed the influence of Anchor Text which originates from within the same domain, using the 'external' Anchor Text: the aggregated Anchor Text from all pages that link to a page in the same domain as the target page. Another aspect of Anchor Text is its development over time: often single snapshots of sites are used to extract links and Anchor Text, neglecting historical trends. Dai and Davison [74] determined Anchor Text importance by differentiating pages' inlink context and creation rates over time. They concluded that ranking performance is improved by differentiating pages with different in-link creation rates, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

In the preceding works, the Anchor Text of a page has been considered as a resource that is complementary to the page content, but treated as two independent representations. Dou et al. [85], Kleinberg [117] took the relationship between source and Anchor Text into

account. Their model distinguished between links from the same website and links from related sites to better estimate the importance of Anchor Text. Similarly, Metzler et al. [132] has overcome the problem of Anchor Text sparsity by smoothing the influence of Anchor Text originating from within the same domain by using 'external' Anchor Text: the aggregated Anchor Text from all pages that link to a page in the same domain as the page to be enriched. In the context of Web archiving, link evidence and Anchor Text could be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [115] computed lexical signatures of lost webpages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost webpages.

So far, we have described works that studied the structure of the Web and how the link structure analysis was exploited for improving retrieval effectiveness. However, all of them focused on using single snapshot of archived websites. Now, we summarize studies that focused on the Web evolution by studying the link development over time. Web link structure is very dynamic and grows following a power law [123]. In the IR community, several works used the temporal information of archived material to improve search effectiveness. Li and Croft [124] proposed a time-based language model based on studying the correlation between time and relevance. Based on the heuristic that the probability of a document being relevant is higher for the most recent documents, they boosted the relevance of recent documents. Jones and Diaz [109] exploited the distribution of document versions over the timeline as an indication of the interval of time relevant to a query. Elsas and Dumais [87] found that documents that are more dynamic over time tend to be more relevant. Finally, Dai and Davison [74] quantified Anchor Text importance by differentiating pages with different incoming link creation rate over time and different historical incoming link context. They concluded that incorporating the importance of Anchor Text over time in the ranking model improves the performance, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

Costa et al. [71] improved the effectiveness of searching Web archives by incorporating temporal features such as number of versions available for the document in the archive, and life span between first and last version of the document. They studied the relation between Web document persistence and relevance. They presented an approach that learns and combines multiple ranking models specific for each period of time based on their believe that a single generic ranking model cannot predict the variance of Web characteristics over a long period of time. They work on a test collection constructed from the Portuguese Web Archive (*PWA*) in order to be used as ground truth for Web Archive Information Retrieval (WAIR) research [69].

The dataset is publicly available at [1], including *269,801* assessed Web document versions. The assessed documents were returned by different ranking models in response to 50 navigational queries. Queries were randomly sampled from the PWA's query log. The PWA consists of archived documents from the Portuguese Web in the period from 1996 to 2009. They found that there is no correlation between lifespan and number of versions, but both are correlated with the relevance of documents. They found that 36% of documents have a life span less than one year; notice that this percentage is different from the percentage found by [136] which is 80%.

Kanhabua and Nejdl [111] studied the evolution of Anchor Text extracted from edit history of Wikipedia. First, they identified a set of entities using the approach introduced by Bunescu and Pasca [59], for each Wikipedia snapshot. The snapshots were generated by partitioning revisions of Wikipedia pages based on one-month granularity. Then, they generate a set of entity-anchor relationships, based on the Anchor Text derived from links pointing to the entities. They found that Anchor Text with temporal information can be candidates for capturing and tracing entities evolution.

Brewington and Cybenko [53] studied the Web pages rate of change. They used the *last-modified* timestamp and the downloading time of Web pages to collect those that are observed over an average of *37* days. Koehler [118] claims that a collection of Web pages tends to stabilize once it reaches a considerable age. They performed accessibility test on a collection of *361 URLs* randomly selected from a Web crawl during a period of *4 years* between December 1996 and February 2001 crawled on weekly basis. Cho and Garcia-Molina [66] proposed estimators for the frequency of change of Web pages by counting the number of accessible days of each Web page. They collected a daily collection of *720,000* pages from *270* popular sites during a period of four months. Fetterly et al. [89] studied the frequency and degree of change of Web pages, they found that the average degree of change varies widely across top-level domains, and the larger pages change more often than smaller pages. They observed that a significant amount of changes on the Web consists of small modifications. Their collection was collected by weekly crawling *150* million URLs , spanning 11 weeks time-interval in 2002. Ntoulas et al. [136] collected Web pages from *154* popular sites gathered from Google Directory. The Web pages were crawled on weekly basis in one year. They observed a high birth and death rates of Web pages and higher turnover rate for hyperlinks. They also observed that most pages that persist over time exhibit only minor changes in their content. Gomes and Silva [93] studied the persistence of both the URLs and the Web page content. They found that most URLs have a short life, and minor fraction of Web pages persist for long period of time. Bordino et al. [52] performed a statistical analysis on a time-aware graph ob-

tained by crawling 190,000 URLs from the *.uk* domain at monthly basis. The URLs were obtained from the Open directory project [8]. Their goal was to investigate whether the link graph is reliable over time by checking it dependent on the a appearance and disappearance of links, or the crawling settings. More over, they quantified turnover rate of Web pages and links inspired by [136]. However, the two collections are different as they are collected based on different crawling settings. The collection used by [52] consists of 133 million pages and 5 billion links gathered by crawling of a real Web domain. while, the collection studied in [136] is limited because it was collected from 154 sites collected by picking up the top-ranked pages from Google Directory. Although, the two collection were collected based on different crawling settings, the statistical analysis of the birth and death rates of Web pages were aligned.

Dai and Davison [74] determined Anchor Text importance by differentiating pages' inlink context and creation rates over time. They concluded that ranking performance is improved by differentiating pages with different in-link creation rates, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive). Kanhabua and Nejdl [111] studied the evolution of Anchor Text extracted from edit history of Wikipedia. They found that Anchor Text with temporal information can be candidates for capturing and tracing the entity evolution.

The link structure and Anchor Text constructed from the archived pages play an important role in assessing the completeness of Web archives. It is impossible to archive the entire Web due its increasing size and evolving content. Therefore, the archived parts of the Web are incomplete. Web archiving theorists acknowledge that the archived parts of the Web is both incomplete and over complete [58, 129]. It is impossible to crawl the Web in a way that all websites and pages are included, for example the *depth-first* crawling strategy excludes websites not in the seeds list, and the *breadth-first* strategy does not crawl discovered websites in depth. Thus both strategies result in an incomplete crawl. On the other hand, Web archives are over complete, as they do not only contain the raw content but also metadata, such as the MIME-type and the date of the crawling time. More over, information that can be constructed from the archived pages, for example, the link structure and Anchor Text. The wealth of information available in the Web archives has been discussed in [142]. Links and Anchor Text can be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [115] computed lexical signatures of lost webpages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost webpages. The use of the link structure and Anchor Text to uncover and reconstruct target pages that were not archived was studied in [106], based on a *depth-*

first crawl of manually selected websites. They used the link structure extracted from archived Web pages to uncover target URLs that were not archived. Links extracted from the archived pages contain evidence of the existence of unarchived target URLs. Based on the link evidence, Huurdeman et al. found that the number of unarchived Web pages is roughly as high as the number of the archived Web pages. Then, they used link evidence to reconstruct basic representations of target URLs. This evidence includes the aggregated Anchor Text, crawl date, and source URLs.

Part I

ACCESSIBILITY OF WEB ARCHIVE CONTENT ALONG THE TIME AXIS – STUDYING A LARGE-SCALE WEB ARCHIVE COLLECTION

3

UNCOVERING THE UNARCHIVED WEB

Web archives preserve the fast changing Web, yet are highly incomplete due to crawling restrictions, crawling depth and frequency, or restrictive selection policies—most of the Web is unarchived and therefore lost to posterity. We propose an approach to recover significant parts of the unarchived Web, by reconstructing descriptions of these pages based on links and anchors in the set of crawled pages, and experiment with this approach on the Dutch Web archive.

Our main findings are threefold. First, the crawled Web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the Web archive. Second, the link and anchor descriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly. Third, the succinct representation is generally rich enough to uniquely identify pages on the unarchived Web: in a known-item search setting we can retrieve these pages within the first ranks on average.

3.1 INTRODUCTION

The advent of the Web has had a revolutionary impact on how we acquire, share and publish information. The vast amount of digital born content is rapidly taking over other forms of publishing, and the overwhelming majority of online publications has no parallel in a material format. Memory and heritage institutions increasingly recognize that such digital born data are as easily deleted as they are published, thereby introducing unprecedented risks to the world's digital cultural heritage [153]. Web archives address this problem by systematically preserving parts of the Web for future generations. It involves a "process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use" [107]. Pioneered by the Internet Archive and later joined by many national libraries, Web archiving initiatives have archived petabytes of Web data. Despite the important attempts to preserve parts of the Web by archiving, a large part of the Web's content is unarchived and hence lost forever. It is impossible to archive the entire Web due to its ever increasing size and rapidly changing content. However, even the parts that have been preserved are incomplete at several levels.

There are two basic strategies for Web archiving, performed by Web crawlers. The first strategy focuses on the automatic harvest-

ing of websites in large quantities (usually a national domain), also known as 'breadth-first crawls'. The second strategy is based on a specific selection policy, where the crawler settings are intended to ensure the complete preservation of specific websites, also known as 'deep crawls' [129, 76, 101]. On the one hand, consider a breadth-first crawl intended to harvest a top-level domain of a country such as the Netherlands. Being the fifth largest top-level domain in terms of registered domains [64], such a crawl may take several months to complete. Additionally, since its settings are designed to discover as many new links as possible, the crawl may not preserve all internal pages within hosts. On the other hand, selective archives might capture more deep levels of harvested websites, since they are focused on crawling specific websites. However, a large degree of linked pages will not be preserved, since the applied crawler settings typically exclude encountered links outside the seed list, even if relevant to a country's cultural heritage.

The overall consequence is that our Web archives are highly incomplete, and researchers and other users treating the archive to reflect the Web as it once was, may draw false conclusions due to unarchived content. The main research question of this chapter is: can we recover parts of the unarchived Web? This may seem like a daunting challenge or a mission impossible: how can we go back in time and recover pages that were never preserved? Our approach is to exploit the hyperlinked structure of the Web, and collect evidence of uncrawled pages from the pages that were crawled and are part of the archive. We show empirically that it is possible to recover significant parts of the unarchived Web, by reconstructing descriptions of these pages based on links and anchors in the crawled pages. We refer to the recovered Web documents as the Web archive's *aura*: the Web documents which were not included in the archived collection, but are known to have existed—references to these unarchived Web documents appear in the archived pages.

Specifically, we investigate the following research questions:

RQ1 *Can we uncover and provide representations of unarchived Web pages exploiting references to them from the archived Web pages?*

RQ1.1 *Can we uncover (a fraction) of unarchived Web pages and hostnames from references in the archived Web pages?*

We exploit the link structure of the crawled content to derive evidence of the existence of unarchived pages, and investigate their number of pages and of domains or hostnames.

RQ1.2 *How rich are the representations that can be created for unarchived Web pages?*

We build implicit representations of unarchived Web pages and domains, based on link evidence and Anchor Text, and investigate the richness (or sparseness) of the descriptions in the

Table 1: Number of documents per year

year	number of docs
2009	17,014,067
2010	38,157,308
2011	53,604,464
2012	38,865,673
	147,641,512

number of incoming links and the aggregated Anchor Text, and break this down over unarchived home pages and other pages.

RQ1.3 *Are the resulting derived representations of unarchived pages sufficient to make them retrievable among millions of archived pages?*

As a critical test, we study the effectiveness of the derived representations of unarchived home pages and deep pages in a known-item search setting. Only if the derived representation characterizes the unique page's content, we have a chance to retrieve the page within the first ranks.

3.2 EXPERIMENTAL SETUP

This section describes our experimental setup: the approach, the dataset, the link extraction methods and the way the links were aggregated for analysis.

3.2.1 *Data*

This study uses data from the Dutch Web archive at the National Library of the Netherlands (KB). The KB currently archives a preselected (seed) list of more than 5,000 websites [140]. Websites for preservation are selected by the library based on categories related to Dutch historical, social and cultural heritage. Each website in the seed list has been categorized using a UNESCO classification code.

Our snapshot of the Dutch Web archive consists of 76,828 ARC files, which contain aggregated Web content. A total number of 148M documents has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data (see Table 1). Basic harvest metadata is available (crawl dates, page modification dates, etc.). Additional metadata is available in separate documentation, which includes the KB's selection list, dates of selection and the manually assigned UNESCO codes by the curators of the KB. In our study, we focus on the documents crawled in 2012.

In our extraction, we differentiate between four different types of URLs found in the Dutch Web archive:

1. URLs that have been archived intentionally as they are included in the seedlist,
2. URLs that have been unintentionally archived due to the crawler's configuration,
3. unarchived URLs, of which the parent domain is included in the seedlist, and
4. unarchived URLs, which do not have a parent domain that is on the seedlist.

3.2.2 *Link Extraction*

We created our dataset by implementing a specific processing pipeline. This pipeline uses Hadoop MapReduce and Apache Pig for data extraction and processing. The first MapReduce job processed all archived webpages contained in the archive's ARC files, and used JSoup to extract links from their contents. For each link, the source URL, target URL, crawldate, Anchor Text and (MD5) hashcode of the source page were kept. Subsequently, this file was matched against the KB's list of seed domains and assigned UNESCO codes, to create a set with an indication if a specific URL is on the seedlist at the moment of crawling, and if it has a UNESCO classification code. A second MapReduce job built a temporary index of all URLs (with their associated crawldate) that occur in the Dutch Web archive, allowing lookups to validate if a given URL exists in the archive or not. Subsequently, the processed files have been joined to create the following list:
(sourceURL, sourceUnesco, sourceInSeedProperty, targetURL, targetUnesco, targetInSeedProperty, anchorText, crawlDate, targetInArchiveProperty, sourceHash)

In our study, we look at the content per year. Therefore, additional steps in our data preparation included deduplication of links per year, to correct for different harvesting frequencies of sites in the archive. While some sites are harvested yearly, other sites are captured biannually, quarterly or even daily. This could result in a large number of links from duplicate pages. To prevent this from influencing our dataset, we deduplicated the links based on their values for year, Anchor Text, source, target, and (MD5) hashcode. The hashcode is a unique value representing a page's content, and is used to detect if a source has changed between crawls. We keep only links to the same target URLs if it originates from a unique source URL.

In our dataset, we include both inter-server links, which are links between different servers (external links), and intra-server links,

which occur within a server (site internal links). We also performed basic data cleaning and processing: removing non-alphanumerical characters from the Anchor Text, converting the source and target URLs to the canonicalized SURTURL format, removing double and trailing slashes, and removing *http(s)* prefixes (see `http://crawler.archive.org/apidocs/org/archive/util/SURT.html`).

3.2.3 *Link Aggregation*

Our next step consisted of aggregating the extracted links by target URL, retaining the captured metadata. In this process, we create a representation that includes the target URL and properties, and grouped data elements with source URLs, Anchor Text and other associated properties. Using another Apache Pig script, we counted different elements, for example the unique source sites and hosts, unique anchor words, and the number of links from seed and non-seed source URLs. We also split each URL to obtain separate fields for TLD, domain, host and filetype. To retrieve correct values for the TLD field, we matched the TLD extension from the URL with a list of all TLDs, while we matched extracted filetype extensions of each URL with a list of common Web file formats.

This aggregated representation containing target URLs, source properties and value counts was subsequently inserted into a MySQL database (13M rows), to provide easier access for analysis.

3.3 EXPANDING THE WEB ARCHIVE

In this section, we study **RQ1.1** *Can we uncover (a fraction) of unarchived Web pages and hostnames from references in the archived Web pages?* We investigate the contents of the Dutch Web archive and quantify the unarchived material that can be uncovered via the archive. Our finding is that the crawled Web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the Web archive.

3.3.1 *Archived Content*

We begin by introducing the actual archived content of the Dutch Web archive in 2012, before characterizing the unarchived contents in the next subsection. Here, we look at the unique text-based webpages (based on MD5 hash) in the archive, totaling in 11,041,113 pages. Of these pages,
10,158,586 were crawled in 2012 as part of the KB's seedlist (92%). An additional 882,527 pages are not in the seedlist but included in the archive (see Table 2). Each 'deep' crawl of a website included in the seedlist also results in additional ('out of scope') material being

Table 2: Unique archived pages (2012)

	on seedlist	%	not on seedlist	%	total
pages	10,158,586	92.0	882,527	8.0	11,041,113

Table 3: Unique archived hosts, domains & TLDs

	on seedlist	%	not on seedlist	%	total
hosts	6,157	14.2	37,166	85.8	43,323
domains	3,413	10.1	30,367	89.9	33,780
TLDs[1]	16	8.8	181	100	181

Table 4: Coverage in archive

mean page count	on seedlist	not on seedlist
per host	1,650	24
per domain	2,976	29
per TLD	634,912	4,876

harvested, due to crawler settings. For example, to correctly include all embedded elements of a certain page, the crawler might need to harvest pages beyond the predefined seed domains. These unintentionally archived contents amount to 8% of the full Web archive in 2012.

We can take a closer look at the contents of the archive by calculating the diversity of hosts, domains and TLDs contained in it. Table 3 summarizes these numbers, in which the selection-based policy of the Dutch KB is reflected. The number of hosts and domains is indicative of the 3,876 selected websites on the seedlist in the beginning of 2012: there are 6,157 unique hosts (e.g. *papierenman.blogspot.com*) and 3,413 unique domains (e.g. *okkn.nl*).

The unintentionally archived items reflect a much larger variety of hostnames and domains than the items from the seedlist, accounting for 37,166 unique hosts (85.8%), and 30,367 unique domains (89.9% of all domains). The higher diversity of the non-seedlist items also results in a lower coverage in terms of number of archived pages per domain and per host (see Table 4). The mean number of pages per domain is 2,976 for the sites included in the seedlist, while the average number of pages for the items outside of the seedlist is only 29.

According to the KB's selection policies, sites that have value for Dutch cultural heritage are included in the archive. A more precise indication of the categories of websites on the seedlist can be obtained by looking at their assigned UNESCO classification codes.

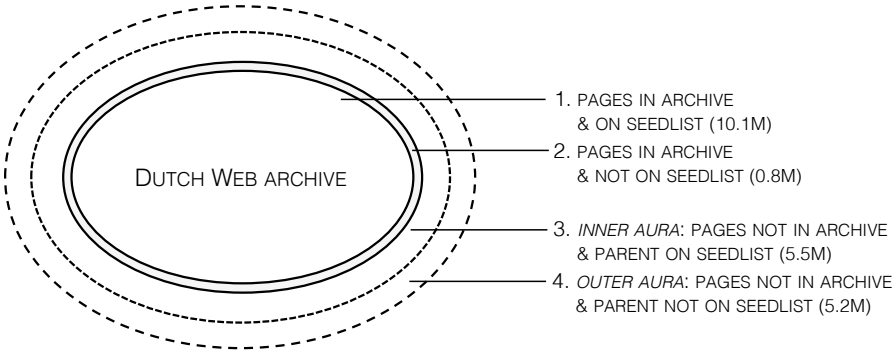


Figure 5: 'Layers' of contents of the Dutch Web Archive (2012)

Table 5: Unarchived *aura* unique pages (2012)

	inner aura	%	outer aura	%	Total
pages	5,505,975	51.5	5,191,515	48.5	10,697,490

In the archive, the main categories are Art and Architecture (1.3M harvested pages), History and Biography (1.2M pages) and Law and Government Administration (0.9M pages). For the sites harvested outside of the selection lists, no UNESCO codes have been assigned. A manual inspection of the top 10 domains in this category (35% of all unintentionally harvested pages) shows that these are heterogeneous: 3 sites are related to Dutch cultural heritage, 2 are international social networks, 2 sites are related to the European Commission and 3 are various other international sites.

3.3.2 *Unarchived Content*

To uncover the unarchived material, we used the link evidence and structure of crawled contents of the Dutch Web archive. We refer to these contents as the Web archive's *aura*: the pages that are not in the archive, but which existence can be derived from evidence in the archive.

The unarchived *aura* has a substantial size: there are 11M unique pages in the archive, but we have evidence of 10.7M additional link targets that are not in the archive. In the following sections, we will focus on this aura, and differentiate between the *inner aura* (unarchived pages of which the parent domain is on the seedlist) and the *outer aura* (unarchived pages of which the parent domain is not on the seedlist). The inner aura has 5.5M (51.5%) unique link targets, while the outer aura has 5.2M (48.5%) unique target pages (see Figure 5 and Table 5).

Table 6: Unarchived unique hosts, domains & TLDs

	inner aura	%	outer aura	%	total
hosts	9,039	1.8	481,797	98.2	490,836
domains	3,019	0.8	369,721	99.2	372,740
TLDs	17	6.6	259	100	259

Table 7: Unarchived *aura* coverage (2012)

mean page count	inner aura	outer aura
per host	609	10
per domain	1,823	14
per TLD	323,881	20,044

Table 8: Unarchived *aura* filetypes

inner aura	count	%	outer aura	count	%
http	4,281,750	77.77	http	3,721,059	71.68
html	351,940	6.39	php	585,024	11.27
php	321,095	5.83	html	582,043	11.21
asp	38,0964	6.92	asp	181,963	3.51
pdf	70,371	1.28	jpg	30,205	0.58

Like the number of pages, also the number of unique unarchived hosts is quite substantial: while *in* the archive there are 43,323 unique hosts, we can reveal a total number of 490,836 hosts in the unarchived aura. There is also a considerable number of unique domains and TLDs in the unarchived contents (see Table 6).

The tables above also show the difference between the *inner* and *outer* aura. The outer aura has a much larger variety of hosts, domains and TLDs compared to the inner aura (Table 6). On the other hand, the coverage in terms of the mean number of pages per host, domain and TLD is much greater in the inner aura than the outer aura (see Table 7). This can be explained by the fact that the pages in the inner aura are closely related to the smaller set included in Web archive's seedlist, since they have a parent domain which is on the seedlist.

Finally, to get an overview of the nature of the unarchived resources, we have matched the link targets with a list of common Web file extensions. From this data, we can derive that the majority of references to the unarchived aura points to textual Web content. Table 8 shows the filetype distribution: the majority consists of URLs without an extension (http), html, asp and php pages for both the inner and outer aura. Only a minority of references are other formats, like pdfs and non-textual contents (e.g. jpg files in the outer aura).

Table 9: TLD distribution

inner aura	count	%	outer aura	count	%
1 nl	5,268,772	95.7	1 com	1,803,106	34.7
2 com	130,465	2.4	2 nl	1,613,739	31.1
3 org	52,309	1.0	3 jp	941,045	18.1
4 net	44,348	0.8	4 org	243,947	4.7
5 int	8,127	0.2	5 net	99,378	1.9
6 *other*	1,954	<0.1	6 eu	80,417	1.6
			7 uk	58,228	1.1
			8 de	44,564	0.9
			9 be	43,609	0.8
			10 edu	29,958	0.6

3.3.3 *Characterizing the "Aura"*

Here, we characterize unarchived contents of the archive based on the top-level domain distribution and the domain coverage.

From the top-level domains (TLDs) we derive the origins of the unarchived pages surrounding the Dutch Web archive. Table 9 shows that the majority of unarchived pages in the inner aura (95.69%) have Dutch origins. The degree of .nl domains in the outer aura is lower, albeit still considerable, with 31.08% of all 1.8M pages. The distribution of TLDs in the outer aura seems to resemble the TLD distribution of the open Web. Even though the regional focus of the selection policy of the Dutch Web archive is apparent in the distribution of the top 10, the comparison does provide indications that the outer aura is more comparable to the full Web. The prominence of the .jp TLD can be explained by the fact that some Japanese social networks are included in the unintentionally harvested pages of the Dutch archive.

Another way to characterize the unarchived contents of the Dutch Web is by studying the distribution of the target domain names. This distribution is quite distinct in the two subsets of the aura: while the inner aura contains many specific Dutch sites, as selected by the KB (e.g. *noord-hollandsarchief.nl* and *archievenwo2.nl*), the outer aura contains a much more varied selection of sites, which include both popular international and Dutch sites (e.g. *facebook.com* and *hyves.nl*), and very specific Dutch sites potentially related to Dutch heritage (e.g *badmintoncentraal.nl*).

To get more insights into the degree of popular sites in the unarchived aura, we compare the domains occurring in the aura against publicly available statistics of websites' popularity. Alexa, a provider of free Web metrics, publishes online lists of the top 500 ranking sites per country, on the basis of traffic information. Via the Internet Archive, we retrieved a contemporary Alexa top 500 list for

Table 10: Coverage of most popular Dutch sites (*Alexa position*)

inner aura	count	outer aura	count
nu.nl (*6*)	74.2K	twitter.com (*9*)	266.7K
wikipedia.org (*8*)	17.4K	facebook.com (*3*)	227.0K
blogspot.com (*15*)	3.5K	linkedin.com (*7*)	184.9K
kvk.nl (*90*)	2.2K	hyves.nl (*11*)	125.6K
anwb.nl (*83*)	1.7K	google.com (*2*)	106.4K

sites in the Netherlands (specifically, `http://web.archive.org/web/20110923151640/alexa.com/topsites/countries/NL`). We counted the number of sites in Alexa's top 100 that occur in the inner and outer aura of the Dutch archive (summarized in Table 10). The inner aura covers 7 sites of the top 100 Alexa sites (including Dutch news aggregator *nu.nl* and *wikipedia.org*), while the outer aura covers as much as 90 of the top 100 Alexa sites, with a considerable number of unique target pages. For these 90 sites, we have in total 1,227,690 URL references, which is 23.65% of all unarchived URLs in the outer aura of the archive. This means that we have potentially many representations of the most popular websites in the Netherlands, even though they have not been captured in the selection-based archive itself.

Summarizing, in this section we have quantified the size and diversity of the unarchived sites surrounding the selection-based Dutch Web archive. We found it to be substantial, with almost as many references to unarchived URLs as pages in the archive. These sites complement the sites collected based on the selection policies, and provide context from the Web at large, including the most popular sites in the country. The answer to our first research question is resoundingly positive: the indirect evidence of lost Web pages holds the potential to significantly expand the coverage of the Web archive. However, the resulting Web page representations are different in nature from the usual representations based on Web page content. We will characterize the Web page representations based on derived descriptions in the next section.

3.4 REPRESENTATIONS OF UNARCHIVED CONTENT

In this section, we study **RQ1.2** *How rich are the representations that can be created for unarchived Web pages?* We build implicit representations of unarchived Web pages and domains, based on link evidence and Anchor Text, and investigate the richness (or sparseness) of the resulting descriptions in the number of incoming links and the aggregated Anchor Text, and break this down over unarchived home pages and other pages. Our finding is that the link and anchor de-

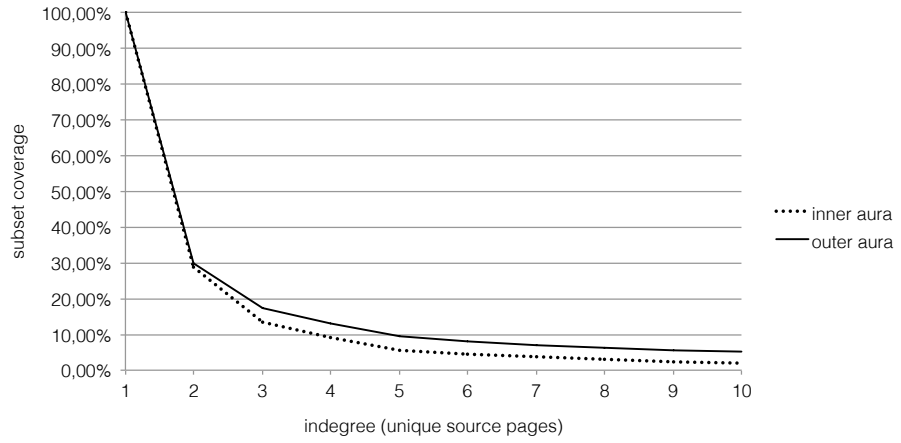


Figure 6: Number of unique source pages (based on MD5 hash) compared to subset coverage

Table 11: Link types

	inner aura	%	outer aura	%
intra-server	5,198,479	94.4	2,065,186	39.8
inter-server	289,412	5.3	3,098,399	59.7
inter & intra-server	18,084	0.4	27,930	0.5

scriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly.

3.4.1 *Indegree*

In general, the representation of a target page is richer if it includes Anchor Text contributed from a wider range of source sites, i.e. has a higher indegree. Therefore, we looked at the number of incoming links for each target URL in our uncovered archive. This is shown in Figure 6, which shows a highly skewed distribution: all target representations in the outer aura have at least 1 source link, 18% of the collection of target URLs has at least 3 incoming links, and 10% has 5 links or more. The pages in the inner aura have a lower number of incoming links than the pages in the outer aura. To check whether this is related to a higher number of intra-server (internal site) links, we also assessed the types of incoming links.

We differentiate between two link types that can be extracted from archived Web content: intra-server links, pointing to the pages in the same domain of a site, and inter-server links, that point to other websites. Table 11 shows the distribution of these types of links of the uncovered aura. It shows that the inner aura has a majority of links from the same source server (i.e. a site on the seedlist), while the outer aura has a much smaller degree of intra-server links. There

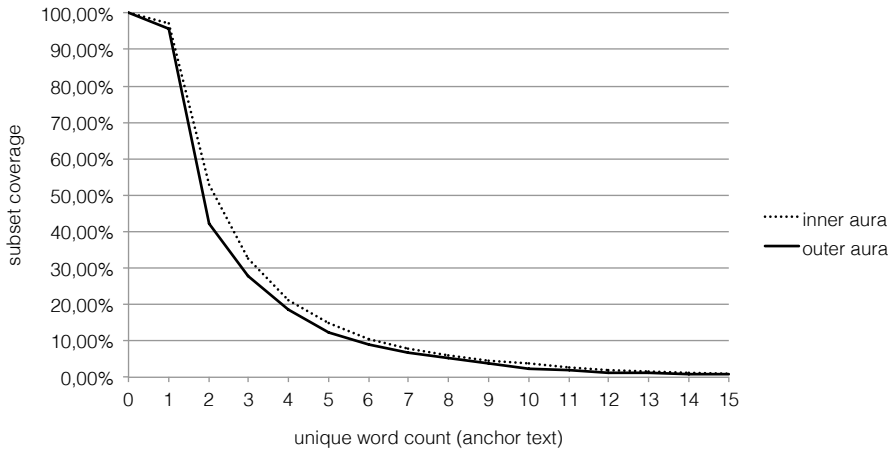


Figure 7: Number of unique words compared to subset coverage

are very few link targets with both intra-server and inter-server link sources in the inner and outer aura.

3.4.2 *Anchor Text Representations*

An influence on the utility of possible representations of sites is also richness of the Anchor Text. In the aggregated Anchor Text representations, we counted the number of unique words in the Anchor Text. Figure 7 shows the number of unique words compared to subset coverage. Like the previous distribution of incoming source links, the distribution of unique Anchor Text is rather skewed. While 95% of all target URLs in the archive have at least 1 word describing them, 30% have at least 3 words as a combined description, and around 3% have 10 words or more (though still amounting to 322,245 unique pages). The number of unique words per target is similar for both the inner and outer aura.

3.4.3 *Homepage Representations*

As mentioned in section 2.3, anchors have been used for homepage finding, since links often refer to homepages. To verify to what extent our dataset contains homepages, we looked at whether a homepage is available for each captured host in the outer aura. We calculated this number by counting the slashes in the target URLs, keeping the pages with a slashcount of 0, and by creating a set of manual filters for homepages (e.g. URLs that contain 'index.html') for pages with slashcount higher than 0. The results of this analysis indicate that for a total of 481,797 hosts, actually 336,387 homepages are available. In other words, 69.8% of all hosts have their (likely) homepage captured in our dataset. This can be important from a preservation and research perspective, since homepages are essential elements of web-

Table 12: Target structure distribution

slashcnt	inner aura	%	slashcnt	outer aura	%
0	3,765	0.1	0	324,782	6.3
1	373,070	6.8	1	921,719	17.8
2	587,416	10.7	**2**	**1,543,129**	**29.7**
3	662,573	12.0	3	535,293	10.3
4	**1,098,947**	**20.0**	4	417,361	8.1
5	535,564	9.7	5	284,237	5.5

sites, but also for the representations that we can generate from the link evidence, because homepages often have a higher indegree and more available Anchor Text.

To obtain a better view of the distribution of pages at different site depths, we also looked at the slashcount of the absolute URLs (see Table 12). From this analysis, we can see that the pages in the outer aura are mainly located at the first levels of the site (i.e. homepage to third level). The links towards the inner aura are pointing to pages that are deeper in the hierarchy, probably because 94% of this subset consists of intra-site link targets (links within a site).

3.4.4 *Qualitative Analysis*

Finally, we provide some examples of representations that we can create for target URLs in this dataset. We start with a homepage with a high indegree from our evaluation sample: *vakcentrum.nl*, a Dutch site for independent professionals in the retail sector. It has 142 inlinks from 6 unique hosts (6 different Anchor Text strings), resulting in 14 unique words. In Table 13 (A) 9 of the unique words (excluding stopwords) are displayed. They provide a basic understanding of what the site is about: a branch organization for independent retailers in the food sector.

For other non-homepage URLs it is harder to represent their contents based on the Anchor Text alone. Take for example *knack.be/nieuws/boeken/blogs/benno-barnard*, a page that is not available on the live web anymore. It only has 2 Anchor Text words: 'Benno' and 'Barnard'. From the URL, however, we can further characterize the page: it is related to news ('nieuws'), books ('boeken') and possibly is a blog. Hence, we have discovered a 'lost' URL, of which we can get an (albeit basic) description by combining evidence. Of course, this varies for each recovered target URL[2], but based on the number of unique words in both Anchor Text and URL, we can get an estimate of the utility of the representation.

2 e.g. *facebook.com/filmhuisbussum* has only few URL words and as Anchor Text 'facebook'

Table 13: Sample aggregated Anchor Text words

(A) vakcentrum [domain]	(B) nesomexico [non-domain]
vakcentrum.nl (6)	mexico (3)
detailhandel (2)	government (1)
zelfstandige (2)	overheid (1)
ondernemers (2)	mexican (1)
levensmiddelen (2)	mexicaanse (1)
brancheorganisatie (1)	beurzen (1)
httpwwwvakcentrumnl (1)	nesomexico (1)
vgl (1)	scholarship (1)
vereniging (1)	programmes (1)

Other pages have a richer description, even if the source links only originate from one unique host. For example *nesomexico.org/dutch-students/study-in-mexico/study-grants-and-loans* is a page that is not available via the live web anymore (3 incomplete captures are located in the Internet Archive). The Anchor Text, originating from *utwente.nl* (a Dutch University website), has 10 unique words, contributed from 2 unique anchors. In Table 13 the combined anchor and URL words are shown, providing an indication of the page's content.

Summarizing, the inspection of the richness of representations of unarchived URLs indicates that the incoming links and the number of unique Anchor Text words have a highly skewed distribution: for few pages we have many descriptions which provide a reasonable number of anchors and unique terms, while the opposite holds true for the overwhelming majority of pages. The succinct representations of unarchived Web pages are indeed very different in nature. The answer to our second research question is mixed. Although establishing their existence is an important result in itself, this raises doubts whether the representations are rich enough to characterize the page's content. We decide to investigate this in the next section.

3.5 FINDING UNARCHIVED PAGES

In this section, we study **RQ1.3** *Are the resulting derived representations of unarchived pages sufficient to make them retrievable among millions of archived pages?* We focus on the retrieval of unarchived Web pages based on their derived representations in a known-item search setting. Our finding is that the succinct representation is generally rich enough to identify pages on the unarchived Web: in a known-item search setting we can retrieve these pages within the first ranks on average.

3.5.1 *Evaluation Setup*

To evaluate the utility of uncovered evidence of the unarchived Web, we indexed 5.19M representations that are in the *outer aura* of the unarchived Web archive contents. These representations consist of a unique assigned ID, the unarchived URL and aggregated Anchor Text of the pages in the outer aura. We indexed these documents using the Terrier 3.5 IR Platform [138], utilizing basic stopword filtering and Porter stemming. Three indexes were created. The first index uses only the aggregated anchor words (*anchT*). We also created a second index (*urlW*), which uses other evidence: the words contained in the URL. Non-alphanumerical characters were removed from the URLs and the remaining words of 20 characters or less were indexed. The third index consists of both aggregated Anchor Text and URL words (*anchTUrlW*).

To create known-item queries, a stratified sample of the dataset was taken, consisting of 500 random non-homepage URLs, and 500 random homepages. Here, we define a non-homepage URL as having a slashcount of 1 or more, and a homepage URL as having a slashcount of 0. These URLs were checked against the Internet Archive (pages archived in 2012). If no snapshot was available in the Internet Archive (for example because of a *robots.txt* exclusion), the URL was checked against the live Web. If no page evidence could be consulted, the next URL in the list was chosen, until a total of 150 queries per category was reached. The consulted pages were used by two annotators to create known-item queries. Specifically, after looking at the target page, the tab or window is closed and the topic creator writes down the query that he or she would use for refinding the target page with a standard search engine. Hence the query was based on their recollection of the page's content, and the annotators were completely unaware of the Anchor Text representation (derived from pages linking to the target). As it turned out, the topic creators used 5-7 words queries for both homepages and non-homepages. The set of queries by the first annotator was used for the evaluation (n=300), the set of queries by the second annotator was used to verify the results (n=100). We found that the difference between the annotators was low: the average difference in resulting MRR scores between the annotators for 100 homepage queries in all indexes was 8%, and the average difference in success rate was 3%.

Subsequently, we ran these 300 queries against the *anchT*, *urlW* and *anchTUrlW* indexes created in Terrier using its default InL2 retrieval model based on DFR, and saved the rank of our URL in the results list. To verify the utility of anchor, URL words and combined repre-

Table 14: Mean Reciprocal Rank (MRR)

MRR	# Queries	anchT	UrlW	anchTUrlW
homepages	150	0.327	0.317	0.489
non-homepages	150	0.254	0.384	0.457
combined	300	0.290	0.351	0.473

sentations, we use the Mean Reciprocal Rank (MRR) for each set of queries against each respective index.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \tag{1}$$

The MRR (1) is a statistical measure that looks at the probability of retrieving correct results. It is the average over the scores of the first correct result for each query (calculated by $\frac{1}{\text{rank}}$). We also compute the success rate at rank 10, that is, for which fraction of the topics do we actually retrieve the correct URL within the first 10 ranks.

3.5.2 *Availability of Pages*

We used unarchived pages uncovered from the Dutch Web archive, that are either available in the Internet Archive, or still available on the live Web, in order to have the ground truth information about the page's content. This potentially introduces bias—there can be some difference between the pages that still are active, or have been archived, and those that are not—but the URLs did not suggest any striking differences. Out of all randomly chosen homepages surveyed, 79.9% were available via either the Internet Archive or the live Web. However, this was not the case for the non-homepages (randomly selected pages with a slash count of 1 or more), as only 49.8% could be retrieved via the Internet Archive or the live Web. The underlying reasons that many URLs could not be archived include restrictive robots.txt policies (e.g. Facebook pages), contents specifically excluded from the archive (e.g. Twitter accounts and tweets), but also links resulting from page scripts (e.g. LinkedIn 'share' buttons). The unavailability of URLs strengthens the potential utility of generated page representations, for example via aggregated Anchor Text, since no page evidence can be retrieved anymore.

3.5.3 *MRR and Success Rate*

MRR scores were calculated for the examined homepages and non-homepages to test to what extent the generated representations suf-

Table 15: Success rates (target page in top 10)

Success@10	# Queries	anchT	UrlW	anchTUrlW
homepages	150	46.7%	39.3%	64.0%
non-homepages	150	34.7%	46.0%	55.3%
combined	300	40.7%	42.7%	59.7%

fice to retrieve unarchived URLs. The final results of the evaluation based on MRR are summarized in Table 14. We found that the MRR scores for the homepages and non-homepages are quite similar, though some differences can be seen. Using the Anchor Text index, the homepages score higher than the non-homepages, possibly because of the richer representations available for these homepages. The scores for the URL words index are naturally higher for the non-homepages: they have longer URLs and therefore more words that could match the words used in the query. Finally, we can see that the combination of anchor and URL words evidence significantly boosts the retrieval effectiveness: the MRR is close to 0.5, meaning that in the average case the correct result is retrieved at the second rank.

We also examined the success rate, that is, for which degree of the topics do we actually retrieve the correct URL within the first 10 ranks? Table 15 shows that again there is some similarity between the homepages and non-homepages. The homepages score better using the Anchor Text index than the non-homepages: 46.7% can be retrieved. On the other hand, the non-homepages fare better than the homepages using the URL words: 46.0% of the non-homepages is included in the first 10 ranks. Again, we see that combining both representations results in a significant increase of the success rate: we can retrieve 64% of the homepages, and 55.3% of the non-homepages in the first 10 ranks.

The MRR scores indicate that Anchor Text in combination with tokenized URL words can be discriminative enough to do known-item search: the correct results can usually be retrieved within the first ranks. Secondly, the success rates show that by combining Anchor Text and URL word evidence, 64% of the homepages, and 55.3% of the deeper pages can be retrieved. This provides positive evidence for the utility of these representations.

The performance on the derived representations is comparable to the performance on regular representations of webpages [97]. Here we used a standard retrieval model, without including various priors tailored to the task at hand [120].

Table 16: Division based on indegree of unique hosts

indegree	pages	word count	MRR anchT	homepage
1	251	2.9	0.29	42.6%
2	28	3.8	0.19	82.1%
3	12	4.5	0.29	100%
4+	9	7.3	0.49	88.9%

3.5.4 *Impact of Indegree*

Another aspect of the evaluation examines the influence of the number of unique inlinks on the richness of Anchor Text representations. For example, the Centre for European Reform (*cert.org.uk*) receives links from 3 unique hosts: *portill.nl*, *europa-nu.nl* and *media.europa-nu.nl*, together contributing 5 unique anchor words, while the page *actionaid.org/kenya* has 1 intra-server link from *actionaid.org*, contributing only 1 anchor word. For the combined 300 topics (domains and non-domains together), we calculated the mean unique word count, the MRR and the degree of homepages in the subset. Table 16 summarizes these results.

It shows that, depending on the number of inlinks from unique hosts, the mean word count rises, but it also illustrates the skewed distribution of our dataset: the majority of pages (251 out of 300) have links from only one source host, while a much smaller set (49 out of 300) have links from 2 or more unique source hosts. The table also provides evidence of the hypothesis that the homepages have more inlinks from unique hosts than non-homepages: at an indegree of 2 or more, the homepages take up more than 80% of the set of pages. We can also observe from the data that the MRR using the Anchor Text index in our sample is highest when having links from at least 4 unique hosts.

Summarizing, we investigated whether the derived representations characterize the unique content of unarchived webpages in a meaningful way. We conducted a critical test cast as a known-item finding task, requiring to locate unique pages amongst millions of other pages—a true needle-in-a-haystack task. The outcome is clearly positive: with MRR scores of about 0.5, we find the relevant pages at the second rank on average, and for the majority of pages the relevant page is in the top 10 results. The answer to our third research question is again positive: we can reconstruct representations of unarchived webpages that characterize their content in a meaningful way.

3.6 DISCUSSION AND CONCLUSIONS

In this study, we proposed a method for deriving representations for unarchived content, by using features extracted from a dataset of archived webpages. We used link evidence to firstly *uncover* target URLs outside the archive, and secondly to *reconstruct* basic representations of target URLs outside the archive. This evidence includes aggregated Anchor Text, source URLs, assigned classification codes, crawl dates, and other extractable properties. Hence, we derived representations of URLs that are not archived, and which otherwise would have been lost.

We tested our methods on the data of the selection-based Dutch Web archive in 2012. The analysis presented above first characterized the contents of the Dutch Web Archive, from which the representations of unarchived pages were subsequently uncovered, reconstructed and evaluated. The archive contains almost as many mentions of unarchived pages as the number of the actually archived pages. Hence, using data extracted from archived pages, information can be recovered about unarchived pages which once closely interlinked with the pages in the archive.

The recovery of the unarchived pages surrounding the Web archive, which we called the 'aura' of the archive, can be used for assessing the completeness of the archive, and may help to extend the seedlist of the crawlers of selection-based archives. Additionally, representations of pages could also be used to enrich the index and provide additional search functionalities. Including the representations of pages in the outer aura, for example, is of special interest as it contains evidence to the existence of top websites that are excluded from archiving, such as Facebook and Twitter. This is supported by the fact that only two years since the data was crawled, 20.1% of the found unarchived homepages and 45.4% of the non-home pages could no longer be found on the live Web nor the Internet Archive.

The evaluation of the unarchived pages described in this study shows that the extraction is rather robust, since both unarchived homepages and non-homepages received similar satisfactory MRR average scores. However, there are some limitations to the method described in this study. The first concerns the aggregation of links by year, which may over-generalize timestamps of the unarchived pages and therefore decrease the accuracy of the representation. Second, the recovered representations are rather skewed, hence most of the uncovered pages have relatively sparse representations, while only a small fraction has rich representations. Third, we used data from a selective archive, whose crawler settings privilege select hostnames and are instructed to ignore other encountered sites. This affects the relative distribution of home pages and non-homepages, both in the

archive as well as in the unarchived pages. In future work we will examine the impact of the crawling strategy.

Web archives preserve Web content for posterity, assuming that what is not selected for archiving might be lost forever. This study shows that it is still possible to recover representations of pages that were not selected for archiving. We have developed a method for uncovering evidence of unarchived pages from Web archives, and for reconstructing representations of their past existence based on link and anchors in crawled pages. Our analysis of the Dutch Web archive crawled in 2012 shows that the number of unarchived pages that can be uncovered is as large as the number of the intentionally archived pages. Although the representation of the unarchived pages based on Anchor Text and link structure is skewed (that is, few uncovered pages have very rich representation while the representation of most pages is relatively poor), our analysis shows that Anchor Text and link information suffice to retrieve the unarchive pages within the first two ranks on average. Our initial results are based on straightforward descriptions of pure Anchor Text and URL components and standard ranking models. In follow up research we will examine the effect of including further contextual information, such as the text surrounding the anchors, and advanced retrieval models that optimally weight all different sources of evidence.

4

TEMPORAL ANCHOR TEXT AS PROXY FOR PAST USER QUERIES

Web archives preserve the fast changing web. While we can archive the web pages, the popularity of queries in the past has usually not been preserved. Previous studies have observed the importance of *Anchor Text* for improving the quality of text search, and have shown that *Anchor Text* is similar to real user queries and documents titles. Other studies have shown that documents titles are similar to the real user queries. We propose an approach to reconstruct the past topics of interest to users that would be provided by query log using temporal *Anchor Text*. First, we study the link graph of four years of Web archive in order to show how the target hosts and *Anchor Text* evolve over time. Second, we investigate the importance of *Anchor Text* over time. Our approach is to rank *Anchor Text* based on their popularity in the archive at specific time. Then, we check the importance of the top ranked *Anchor Text* in the public Web at the same time. In order to achieve this, we used the *WikiStats* dataset which aggregates page views of Wikipedia pages. Using exact string matching between top ranked *Anchor Text* and Wikipedia titles in the *WikiStats* dataset, we find a high percentage of overlap (approximately 57%). Our data strengthens the hypothesis that *Anchor Text* may be used as a proxy for actual query volume.

4.1 INTRODUCTION

Despite the important attempts to preserve parts of the web by archiving, a large part of the web's content is unarchived and hence lost forever. In practice it is not feasible to archive the entire web due to its ever increasing size and rapidly changing content. The overall consequence is that our web archives are highly incomplete. On the other hand the Web archive is too complete because it it contains additional information about a Web page, more than its content, such as archived date, outlinks and *Anchor Text*.

Queries that represent the past interests of real users, using the archived Web as it was, are usually not available, because they were not preserved. Motivated by studies which showed that *Anchor Text* is similar to documents titles and real users queries [86, 108], we use the important (popular) *Anchor Text* as proxy for queries in the past. In this chapter, we study how the link graph evolves over time; specifically, we focus on target hosts and *Anchor Text*. We investigate evolution of the *Anchor Text* over time in order to understand what

Table 17: Number of seeds and archived objects over the years

year	# of seeds	# of archived objects
2009	2,491	17,014,067
2010	3,312	38,157,308
2011	3,508	53,604,464
2012	4,085	38,865,673
		147,641,512

was important in Web. In different words, we use *Anchor Text* with their associated timestamps to reconstruct past popular topics. We use topic to refer to user information needs which might consists of one or multiple words.

RQ2 *Can we identify past popular topics using anchor text associated with hyperlinks of the Web archive?*

4.2 SETUP

4.2.1 *Dataset*

This study uses data from the Dutch Web archive at the National Library of the Netherlands (KB). The KB currently archives a pre-selected (seed) set of more than 5,000 websites [140]. Websites for preservation are selected by the library per category related to Dutch historical, social and cultural heritage. Our snapshot of the Dutch Web archive consists of 76,828 ARC files, which contain aggregated web content. Each ARC file contains multiple archived records (content plus response header). A total number of 148M documents has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data. Basic harvest metadata is available (crawl dates, page modification dates, etc.). Additional metadata is available in separate documentation, which includes the KB's selection list, date of selection, and manually assigned UNESCO codes (by curators of the KB). Table 17 summarizes the number of websites added to the selection list and the total number of Web objects archived over the years.

4.2.2 *Link Extraction & Aggregation*

We extract a link structure from the archived objects that have `text/html` as MIME-type. The main percentage (approximately 70%, per year) of the archived web objects are HTML-based textual content. In order to extract the links from the archive, we use MapReduce to process all archived web objects contained in the archive's ARC files.

During processing of the archived objects, JSoup[1] was used to extract anchor links from web objects that have `text/html` as MIME-type. For each found anchor link, we keep the source URL (which is the URL of the page that has the link), target URL (which is the URL of the page that the link is pointing to), and the Anchor Text of the link (a short text describing the target page). The archived pages have meta data of about the archived page such as the crawl date. We combine the year and the month of the crawl date with link information (YYYYMM). In addition to that, we keep the hash code (MD5) of the source page. More precisely, we keep the following information:

```
(sourceURL, targetURL, linkType, anchorText, crawlDate,
                        sourceHash)
```

The link type (linkType) indicates whether the link is internal link or external link. An internal link has the same domain-name for both source and target (intra-domain), while an external link the domain-name of the source URL is different from that of the target URL (an inter-domain link). We use linkType to keep only external links. We partition these links based on one-year, and one-month granularity.

In each partition, we deduplicate the links based on their values for sourceURL, targetURL, Anchor Text, year and a hash of the source's content. Different sites in the seeds list are harvested at different frequencies; while most sites are harvested only once a year, some sites are crawled more frequently. At the end of the pipeline, we keep the following information:

```
(sourceURL, targetURL, anchorText, crawlDate)
```

4.2.3 *Wikipedia Page Views Statistics*

The query log that would provide the topics that are of interest to users in the past is not available. Therefore, we used different source as indicator of past popular topics. Motivated by the studies which showed that Anchor Text is similar to document titles and user queries, we used the *WikiStats* project dataset [135]. The *WikiStats* dataset is an aggregated dataset from the `Page view statistics for Wikimedia projects`[2], which keeps the request history of articles from Wikipedia or from another projects. For each article, it keeps the title and the number of requests. *WikiStats* consists of weekly absolute views for Wikipedia pages in the period from January 2008 and January 2015. This gives the number of page views for the Wikipedia pages, the top-level domain (TLD) of the page (such as NL for the Netherlands), and the page's title. Because our snapshot of the Dutch Web archive covers the period between February 2009 and December

1 `http://jsoup.org/`
2 http://dumps.wikimedia.org/other/pagecounts-raw/

2012, we focused on the same period of the *WikiStats* dataset. We partitioned the dataset in this period based on one-month granularity and one-year granularity, keeping only Wikipedia titles which have more than 1,000 page views.

4.3 ANALYSIS

4.3.1 *Hosts Evolution*

In Section 4.2.2, we introduced our approach of extracting the link graph from the archived *text/html* pages combined with metadata such as the crawl date, generating different partitions at different granularities. In this section study the importance of hosts (sites) in the archive over time.

First, we experiment with partitions based on the year granularity. For each partition, we generate the host of both the source page and target page in each link. For example, the host of `https://www.cwi.nl/research/groups/information-access` URL is *cwi.nl*. Multiple links from the same source host will be considered one, we do that by deduplicating the data based on the source host, target host, and Anchor Text.

After that we aggregate the links by target host. Finally, we rank the target hosts based on the number of incoming links; which corresponds to the number of unique source hosts pointing to them. Table 19 shows the top ranked hosts per year. We observe that the ranks of the top hosts vary over the years. By considering the top 1,000 hosts per year, we find no correlation (using Kendall's τ) between the ranked lists of hosts in different years; the strongest negative correlation τ was -0.982 between 2011 and 2012. Table 18 shows the percentage of new hosts in our crawls over the years, considering different thresholds of the top hosts. Here, a host is considered new in a particular year if it does not appear in any previous year.

Next, we experiment with aggregating links by target host, based on the one-month granularity. Table 20 and Table 21 show the top hosts per month in 2009, illustrating that the top hosts vary over the months as well. The number of target hosts varies per month, with an average of 53,215 hosts per month, where 25% these hosts are new.

4.3.2 *Anchor Text Evolution*

In this section, we look into the usage of Anchor Text over time. For each partition $\mathcal{A}_t$ at a given time granularity, we aggregate links by Anchor Text. The number of links using Anchor Text a represents the frequency of a in partition $\mathcal{A}_t$. We used this relative frequency to represent the importance of Anchor Text a in the archive at spe-

Table 18: Percentage of new target hosts over the years considering the top 1,000, 5,000, and 10,000 hosts.

year	Top 1,000	Top 5,000	Top 10,000
2010	37.5	38.3	38.9
2011	26.8	27.3	27.4
2012	19.1	21.2	21.4
Mean	27.80	28.9	29.2

cific time granularity t (`archive-based popularity`), computing the importance of the Anchor Text as follows:

$$I(a, \mathcal{A}_t) = \frac{f(a, \mathcal{A}_t)}{max_{\mathcal{A}_t}} \tag{2}$$

where $f(a, \mathcal{A}_t)$ is the frequency of Anchor Text a in partition $\mathcal{A}_t$, and $max_{\mathcal{A}_t}$ is the maximum frequency of any Anchor Text in partition $\mathcal{A}_t$.

$$max_{\mathcal{A}_t} = \max_a f(a, \mathcal{A}_t) \tag{3}$$

First, we investigate the evolution of Anchor Text over time. Therefore, for the Anchor Text in partition $\mathcal{A}_t$, we compute the percentage of new Anchor Text at the time of t. An Anchor Text is considered new in $\mathcal{A}_t$ if it does not appear in any previous partition.

$$new(a, t) = \begin{cases} 1, & \text{if } a \notin \bigcup_{i<t} \mathcal{A}_i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{A}_i$ represents any partition with time granularity less than the time granularity of $\mathcal{A}_t$. Based on the partitions of one-year granularity, with an average of 999,695 distinct Anchor Text per year, we find that 59% of Anchor Text are new (average across the percentage of all years). Based on the partitions of one-month granularity, 17,024 links with distinct Anchor Text exist per month. The average percentage of new Anchor Text per months is 34%.

4.3.3 *Matching Anchor Text To Wikipedia Title*

In the related work chapter (Section 2.3), we have discussed a series of studies that showed that document titles are close to real user queries, and that Anchor Text is similar to both document titles and real user queries. We therefore hypothesize that we may be able to reconstruct

Table 19: Top ranked hosts over the years.

2009	2010	2011	2012
vriezenveners.nl	hetutrechtsarchief.nl	wikipedia.org	**twitter.com**
mi-website.es	wikipedia.org	hetutrechtsarchief.nl	hetutrechtsarchief.nl
startpagina.nl	europa-nu.nl	biblion.nl	wikipedia.org
fd.nl	bibe.library.uu.nl	**twitter.com**	europa.eu
z24.nl	**twitter.com**	europa-nu.nl	bibe.library.uu.nl
wikipedia.org	belastingdienst.nl	europa.eu	wordpress.com
blogspot.com	europa.eu	bibe.library.uu.nl	blogspot.com
deviantart.com	vriezenveners.nl	blogspot.com	europa-nu.nl
co.uk	startpagina.nl	co.jp	youtube.com
volkskrant.nl	minszw.nl	youtube.com	vriezenveners.nl
gencircles.com	uva.nl	co.uk	co.uk
sitestat.com	readspeaker.com	wordpress.com	google.com*
belastingdienst.nl	blogspot.com	leidenuniv.nl	leidenuniv.nl
web-log.nl	co.uk	google.com*	ebay.com
startkabel.nl	google.com*	belastingdienst.nl	rijksoverheid.nl
imageshack.us	sitestat.com	startpagina.nl	marktplaats.nl
readspeaker.com	amazon.com	vriezenveners.nl	overheid.nl
google.com*	wordpress.com	amazon.com	co.jp
hva.nl	youtube.com	readspeaker.com	knaw.nl
digischool.nl	ebay.com	ligfiets.net	volkskrant.nl
nrc.nl	omroep.nl	zijpermuseum.nl	nuzakelijk.nl
trouw.nl	volkskrant.nl	co.cc	zie.nl
wordpress.com	web-log.nl	ebay.com	startpagina.nl
photobucket.com	ligfiets.net	kennisnet.nl	facebook.com
ugo.com	nrc.nl	tue.nl	tue.nl

past popular topics based on Anchor Text used in the past. Similar to the use of wikipedia in [157], we used the *WikiStats* dataset (described in Section 4.2.3) in order to find how the important Anchor Text in the archive were related to popular queries in the past on the public Web. We consider the number of page views of Wikipedia titles that match Anchor Text to represent the importance of that Anchor Text in the public Web (`web-based popularity`). We study the similarity between Anchor Text and Wikipedia titles varying temporal granularity. We used exact string matching to match Anchor Text with titles of Wikipedia pages in the *WikiStats* dataset, using the same time granularity. Matching was done after transforming both Anchor Text and Wikipedia titles into lower case. For each partition at time t, we rank the Anchor Text based on archive-based popularity, after which we check at different thresholds k how many of the top-k Anchor Text occurrences in the *WikiStats* dataset (in the partition at time t of the *WikiStats* dataset). Table 22 summarizes the percentage of Anchor Text that have a matched Wikipedia title. As we observe in the Table, a high percentage of the top ranked Anchor Text has a matching Wikipedia title. For example, 56% of the top-1k Anchor Text occurrences in the 2009 partition were found also in the 2009 partition of

Table 20: Top hosts per month. **02-06/2009**

200902	200903	200904	200905	200906
fd.nl	adobe.com	volkskrant.nl	deviantart.com	knnv.nl
z24.nl	ppsi.nl	trouw.nl	imageshack.us	blogspot.com
ugo.com	schoolenveiligheid.nl	nrc.nl	photobucket.com	waarneming.nl
volkskrant.nl	omroep.nl	emancipatie.nl	avs.nl	google.com
volny.cz	minocw.nl	nd.nl	intermediair.nl	europa.eu
digischool.nl	pestweb.nl	szw.nl	intermediairforward.nl	web-log.nl
fateback.com	eu.int	europa-nu.nl	sitestat.com	startpagina.nl
sitestat.com	europa.eu	cgb.nl	independer.nl	volkskrantblog.nl
trouw.nl	aps.nl	mozilla.org	blogspot.com	wordpress.com
aol.com	ez.nl	ad.nl	indymedia.nl	co.uk
nrc.nl	aob.nl	refdag.nl	wikipedia.org	wikipedia.org
wikipedia.org	overheid.nl	mozilla.com	wikimedia.org	decontrabas.com
blogspot.com	kpcgroep.nl	lbr.nl	punt.nl	google.nl
chinesefreewebs.com	kennisnet.nl	wordpress.com	co.uk	vpro.nl
typepad.com	justitie.nl	rotterdamsdagblad.nl	blogger.com	wikimedia.org
szw.nl	telegraaf.nl	parool.nl	wordpress.com	eu.int
pretoriashow.com	volkskrant.nl	wikimedia.org	free.fr	omroep.nl
ad.nl	vfpf.nl	europa.eu	youtube.com	gov.uk
co.uk	havenmuseum.nl	telegraaf.nl	libero.it	americantaskforce.org
freewebtown.com	rutgersnissogroep.nl	cnet.com	aol.com	imageshack.us

Table 21: Top hosts per month. **07-12/2009**

200907	200908	200909	200910	200911	200912
volkskrantblog.nl	seniorweb.nl	readspeaker.com	mi-website.es	vriezenveners.nl	hva.nl
anwb.nl	fietsersbond.nl	belastingdienst.nl	startpagina.nl	gencircles.com	startpagina.nl
wordpress.com	archined.nl	cwi.nl	startkabel.nl	startkabel.nl	blogspot.com
adobe.com	begraafplaats.org	artsennet.nl	fd.nl	startpagina.nl	wikipedia.org
google.com	co.uk	wordpress.com	volkskrant.nl	deviantart.com	google.com
anwbentreebewijs.nl	sitestat.com	w3.org	z24.nl	ugo.com	co.uk
waverunner.nl	drenthe.nl	europa.eu	digischool.nl	readspeaker.com	web-log.nl
wikipedia.org	wikipedia.org	knzb.nl	wikipedia.org	belastingdienst.nl	twitter.com
postbus51.nl	site-id.nl	wikipedia.org	sitestat.com	wikipedia.org	wikimedia.org
pharosreizen.nl	google.com	imageshack.us	trouw.nl	photobucket.com	lexius.nl
live.com	overheid.nl	google.com	web-log.nl	imageshack.us	blogger.com
w3.org	knhb.nl	oreilly.com	nrc.nl	twitter.com	omroep.nl
volkskrant.nl	nai.nl	co.uk	co.uk	youtube.com	wordpress.com
amsterdam.nl	amsterdam.nl	overheid.nl	szw.nl	blogspot.com	creativecommons.org
telekom.at	xs4all.nl	google.nl	members.lycos.co.uk	blogger.com	greenpeace.org
vrom.nl	uitvaartmedia.com	photobucket.com	ifrance.com	avs.nl	hanze.nl
gelderlander.nl	leidenuniv.nl	myspace.com	kennisnet.nl	google.com	technorati.com
google.nl	tudelft.nl	businessweek.com	lycos.nl	co.uk	youtube.com
youtube.com	uitvaartinformatie.nl	uitvaart.nl	ad.nl	wikimedia.org	hszuyd.nl
belastingdienst.nl	volkskrant.nl	blogspot.com	blogspot.com	independer.nl	vpro.nl

Table 22: Absolute count and percentage of Anchor Text per year that has a Wikipedia title match at different thresholds.

	Top-1k		Top-5k		Top-10k	
year	count	%	count	%	count	%
2009	559	55.9	2488	49.8	4259	42.59
2010	585	58.5	2326	46.5	3350	33.50
2011	572	57.2	2466	49.3	3995	39.95
2012	564	56.4	2340	46.8	4186	41.86

the *WikiStats* dataset. We observe that the percentage of overlap between Anchor Text and the *WikiStats* dataset partitions decreases as we increase the threshold of the top-k. The percentage reaches 26% (averages across all partitions) when we consider all Anchor Text in the one-year partition.

Table 23 shows a comma-separated sample of Anchor Text taken from the top-1k popular Anchor Text in 2012 which do not have a match of any Wikipedia titles in 2012 of the *WikiStats* dataset. Some of these are uninformative having a specific purpose, such as *login* to proceed. Some Anchor Text have no match because of limitations due to our approach of looking for exact string match between the Anchor Text and the Wikipedia titles. For example the Anchor Text *filmpje* has no match but in the *WikiStats* dataset there is a page with title *filmpje!*. Likewise, *nunl* has no exact match, however there is a Wikipedia page with title *nu.nl*. In the future, our approach should consider these cases by applying additional pre-processing steps like stemming and stopping, and generalizing from exact match to matches with low edit distance. The list of Anchor Text at the top-1k in 2012 that have a match with Wikipedia title is shown in Table 24. We observe that some of the Anchor Text correspond to cities in the Netherlands such as Amsterdam, Rotterdam, Groningen, Utrecht and Den Haag (all are major cities in the Netherlands). Another category of the top Anchor Text is related to social websites such as twitter, linkedin, flickr, and vimeo. A different category of Anchor Text consists of the major Dutch daily newspapers such as de Volkskrant, Telegraaf, Trouw, and NRC handelsblad. The 'uitzending gemist' occurrence is related to a web service of the Dutch Public Broadcasting (NPO) that offers a free on demand video for nation broadcasts. The 'belastingdienst' Anchor Text is about a governmental service related to the Dutch national tax office.

Based on the one-month granularity, on average 26% of all Anchor Text over all months has an exact match with a Wikipedia title (using all domains). The highest percentage of Wikipedia titles that match the Anchor Text originate from the *NL* domain (around 55%). By ranking the Anchor Text per each one-month granularity based on

Table 23: List of Anchor Text in the top-1k of 2012, that have no matching Wikipedia title.

ga naar website van de fabrikant, word vaste donateur of doneer online via de website van dit goede doel, create your own free blog on wordpresscom, filmpje, vacatures, log in to proceed, wordpresscom, view more information, grotere kaart weergeven, inlichtingen, routebeschrijving, powered by wordpresscom, more information, projectinformatie, volg ons op twitter, nunl, eigen homepage, inschrijven,

Table 24: List of Anchor Text the top-1k of 2012 which have matching Wikipedia titles.

twitter, tweet, linkedin, hyves, jaarverslag, onderzoek, persbericht, pdf, weblog, wordpress, flickr, rapport, rss, vimeo, bron, amsterdam, programma, blogger, de volkskrant, brief, trouw, utrecht, details, samenvatting, rotterdam, groningen, joomla, volkskrant, klik, webwinkel, uitzending gemist, belastingdienst, deel, nrc handelsblad, bericht, den haag, de telegraaf, nrc,

the archive-based popularity, we find that 42.5% of Anchor Text in the top-1k has match with Wikipedia titles.

4.4 CONCLUSIONS AND FUTURE WORK

In this study, we looked into the viability of a new approach of using the evolution of *Anchor Text* over time to reconstruct information that would be similar to real user queries in the past. Our hypothesis is based on studies that have shown that *Anchor Text* behaves similar to both real user queries and documents titles. We used the link structure extracted from the Dutch Web archive to identify the most popular target hosts over time, and to get the most popular *Anchor Text* over time. The link structure was extracted from archived `text/html` archived pages in the Dutch Web archive in the period between February 2009 and December 2012. In order to understand the importance of the *Anchor Text*, we rely on the *WikiStats* dataset, which provides an aggregation of page views of Wikipedia pages. We investigate the exact matches between *Anchor Text* and Wikipedia titles, where both datasets (the link structure and the *WikiStats*) were partitioned based on one-month and one-year granularity. Our analysis of the target hosts shows that target hosts evolve significantly. Based on the one-month granularity, on average 25% among all hosts per month are new. We experiment with finding popular *Anchor Text* per time granularity, by ranking *Anchor Text* based on their popularity in the archive. We find that a high percentage of *Anchor Text* in the top ranks have a match with Wikipedia titles in the *WikiStats* dataset.

Based on the one-year granularity, we found that 57% of the top-1k *Anchor Text* has matching Wikipedia titles. We conclude from our data that the most important *Anchor Text* provides a view of entities in the Netherlands. We cannot however conclude that evolution of *Anchor Text* serves as a proxy for past query logs . There are some limitations that will consider in the future work. First, matching *Anchor Text* and Wikipedia titles analysis, suggests a room for improving our approach by applying additional pre-processing steps like stemming and stopping, and generalizing from exact match to matches with low edit distance. Second, we test our approach on a 'deep crawl' which is based on a few thousands of seeds. In the future, we will test our approach on a 'breadth-first crawl' like the Common Crawl dataset[3].

3 https://commoncrawl.org/

5

COMPARING TOPIC COVERAGE IN BREADTH-FIRST & DEPTH-FIRST CRAWLS

Web archives preserve the fast changing Web by repeatedly crawling its content. The crawling strategy has an influence on the data that is archived. We use link Anchor Text of two Web crawls created with different crawling strategies in order to compare their coverage of past popular topics. One of our crawls was collected by the National Library of the Netherlands (KB) using a *depth-first* strategy on manually selected websites from the *.nl* domain, with the goal to crawl websites as completes as possible. The second crawl was collected by the *Common Crawl* foundation using a *breadth-first* strategy on the entire Web, this strategy focuses on discovering as many links as possible. The two crawls differ in their scope of coverage, while the KB dataset covers mainly the Dutch domain, the *Common Crawl* dataset covers websites from the entire Web. Therefore, we used three different sources to identify topics that were popular on the Web; both at the global level (entire Web) and at the national level (*.nl* domain): Google Trends, *WikiStats*, and queries collected from users of the Dutch historic newspaper archive. The two crawls are different in terms of their size, number of included websites and domains. To allow fair comparison between the two crawls, we created sub-collections from the *Common Crawl* dataset based on the *.nl* domain and the KB seeds. Using simple exact string matching between Anchor Text and popular topics from the three different sources, we found that the *breadth-first* crawl covered more topics than the *depth-first* crawl. Surprisingly, this is not limited to popular topics from the entire Web but also applies to topics that were popular in the *.nl* domain.

5.1 INTRODUCTION

Web archives are created by crawling Web pages following a crawling strategy defined by the institutions. One strategy is to crawl a manually selected set of websites (called the crawler's *seeds*) and to harvest these websites in depth (*depth-first* crawl). Another strategy automatically crawls as many websites as possible (usually the national domains), but not in depth (*breadth-first* crawl). Both crawling strategies result in incomplete crawls, as both strategies exclude websites. *Depth-first* ignores websites outside the seeds list, and *breadth-first* archives websites incompletely as it does not follow the links to sub-pages. On top of the content of websites, Web archives also

preserve information registered by crawlers such as the date of the crawl, the timestamp of the last modification of the page, the MIME-type, and information that can be derived from the archived pages, for example hyperlinks and Anchor Text.

Web archives preserve content which may no longer be available on the Web. We explore how well the collections resulting from different crawling strategies cover content related to topics that were in the focus of Web users in a particular time period. We perform our analysis on two Web archive collections harvested in 2014 using different crawling strategies. The first collection is a crawl from the entire Web harvested by the *Common Crawl* foundation using the *breadth-first* crawling strategy. The second collection is the Dutch Web archive collection preserved by the National Library of The Netherlands[1] (KB). Here, the *depth-first* strategy was applied to manually selected websites (*KB seeds*) related to the Dutch history, social, and culture heritage. We propose to use Anchor Text specified in hyperlinks extracted from the two collections to investigate their coverage of the topics that were of interest to users in the same year (2014). Users of Web search engines express their information needs by issuing queries. User queries collected from major search engines would be the best record of popular topics. However, these queries were not available for us. Therefore, we used different sources as indicators of the trending topics on the Web at the time when the crawls we used were collected (2014). Since our crawls originate from the entire Web (*Common Crawl* crawl) and from the Dutch domain (KB crawl), we looked for popular topics both worldwide and on the national level. Our first source is Google Trends. Google provides a list of the top searched terms on the entire Web, and in the given country domain. The second source is the *WikiStats* which aggregates page views of Wikipedia pages. Again we focus on all Wikipedia pages (in all languages), and the pages written in Dutch. Finally, we use queries collected from users searching the Dutch digital newspaper archive via the KB's Delpher[2] interface. These are three heterogeneous sources, the first and the third are real user queries, the second consists of Wikipedia titles associated with their frequency of views over time. We use these sources to represent users interests, which we refer to as topics. We use topic to refer to user information needs which might consists of one or multiple words.

RQ3 *How does the crawling strategy impact the Web archive's coverage of past popular topics?*

1 www.kb.nl
2 www.delpher.nl

5.2 SETUP

In this section we describe the two crawls on which we base our analysis. Then, we introduce the pipeline of extracting hyperlinks and Anchor Text from the crawls. After that, we discus how we zoom in the link structure of *Common Crawl* dataset to generate subsets based on filters synthesized from the KB dataset in order to allow a fair comparison. Finally, we introduce the sources that we used to identify popular topics.

5.2.1 *Data*

KB dataset

The KB archives a pre-selected set of more than 10,000 websites (*seeds*) with the aim to crawl these websites as complete as possible. The selection is based on categories related to Dutch historical, social and cultural heritage. The websites are categorized by curators of the KB using the *UNESCO* classification code. The crawling frequency varies between yearly, biannually, quarterly, and daily, for example news agency websites (such as nu.nl). Our snapshot of the Dutch Web archive between February 2009 and May 2015 consists of 150,557 files in *ARC*[3] format, which contain aggregated web content. Each ARC file contains multiple Web objects, in total, 251,591,618 objects exist in the ARC files. We focus on data crawled in 2014, as we have only access to *Common Crawl* pages crawled in that year.

Common Crawl *dataset*

Common Crawl[4] is a non-profit organization aiming to build and maintain an openly accessible repository of archived Web crawls. We use the crawl collected in March 2014, which consists of 2.8 billion Web pages.

5.2.2 *Anchor Links Extraction*

From the two datasets, we extracted hyperlinks from the archived objects with *text/html* as MIME-type. For that we used MapReduce to process all archived web objects contained in the archive's ARC files. During the processing of the archived objects, we used JSoup[5] to extract anchor links (*a*) in order to be able to focus on links between textual content. For each anchor link, we kept the URL of the page that contains the link *source*, the URL of the *target*, and the Anchor

3 http://archive.org/web/researcher/ArcFileFormat.php
4 http://commoncrawl.org/
5 http://jsoup.org/

Text specified in the link. Based on the crawl-date, we keep pages crawled in 2014. The Anchor Text pointing to the target pages was used in that year. Depending on the source URL and target URL, the link can be an internal link or external link. An internal link has the same domain-name for both source and target (intra-domain), while for an external link the domain-name of the source URL is different from that of the target URL (an inter-domain link). We limit our analysis to the external links as it is of more interest to look into links between different hosts (sites). By discarding internal links we exclude links from menus and other non-content information. The exact URLs may change frequently, while we are really interested in Anchor Text used by one site to link to another site. Therefore, we replace both the source URL and the target URL by their hosts (site name) before we analyze the data. This pre-processing can be viewed as a process to smooth the graph structure to maintain the most salient information. We deduplicate the links based on their values for source, target, and Anchor Text for KB dataset (*Common Crawl* dataset consists of one crawl). This prevents the differences in crawling frequency to influence our analysis. At the end of this pipeline, we keep (*sourceHost, targetHost, anchorText*). We refer to the links extracted from the KB dataset as $\mathrm{KB}_{\mathtt{links}}$, and links extracted from the *Common Crawl* dataset as $\mathrm{CC}_{\mathtt{links}}$.

5.2.3 *Link Subsets from* Common Crawl

The two crawls differ in terms of size, number of crawled websites and web pages, and the domains of the crawled websites. These differences are reflected in the extracted links structure. The number of links extracted from the *Common Crawl* dataset is 559x times larger than the number of KB links, (see Table 25). Therefore, in addition to performing one-to-one comparison between the two crawls, we generate subsets from the $\mathrm{CC}_{\mathtt{links}}$ by mapping it to the Dutch domain in two different ways: First, we focused on pages that originate from the *.nl* domain. This was done by keeping only links from the $\mathrm{CC}_{\mathtt{links}}$ whose *source hosts* are from the *.nl* domain. We refer to the set as $\mathrm{CC}_{\mathtt{links}} \cap \mathrm{NL}_{\mathtt{tld}}$. Second, the KB crawl is based on a list of manually selected websites (KB seeds). We used the hosts of the KB seeds to generate another subset of links from the $\mathrm{CC}_{\mathtt{links}}$, based on links with *source hosts* from the KB seeds. We refer to this subset as $(\mathrm{CC}_{\mathtt{links}} \cap \mathrm{KB}_{\mathtt{seeds}})$. Finally, we investigate the impact of Anchor Text associated with targets of links in the KB dataset on the topic coverage of the $\mathrm{CC}_{\mathtt{links}}$. In order to do that, we dropped links from $\mathrm{CC}_{\mathtt{links}}$ in which the *target hosts* are targets of links in the $\mathrm{KB}_{\mathtt{links}}$. We refer to this set of filtered links as $(\mathrm{CC}_{\mathtt{links}} \setminus \mathrm{KB}_{\mathtt{targets}})$. These subsets allow us to investigate whether the KB seeds list comprises the part of the Dutch Web that is essential from the perspective of

Table 25: Number of unique links in each dataset.

Links Dataset	Num. of links
KB_{links}	3,033,855
CC_{links}	1,696,102,933
$CC_{links} \cap NL_{tld}$	5,128,501
$CC_{links} \cap KB_{seeds}$	2,629,765
$CC_{links} \setminus KB_{targets}$	1,174,261,413

topic coverage, or whether a broader and less deep crawl would still contain sufficient information.

5.2.4 *Sources of Topics*

Our assumption is that the *Common Crawl* (a *breadth-first* crawl) covers more global topics, and that the KB (a *depth-first* crawl) covers more topics from the *.nl* domain. In order to validate our assumption, we use different sources to identify which topics were popular on the Web, topics that attracted attention in the entire Web (global) and topics that were only picked up in the *.nl* domain.

Google Trends

Google Trends[6] is a public resource, which lists the most searched queries in the global Web or per country in a given year. For our analysis, we use global trends and the trends searched in the Netherlands in 2014 (the year of our crawls).

Wikipedia Page Views Statistics

The *WikiStats* dataset [135] consists of the number of views for Wikipedia pages. The goal is to show how the interest in Wikipedia pages changes over time, and allows comparison between chosen Wikipedia pages. The views are aggregated from the *Page view statistics for Wikimedia projects*[7], which aggregates the request history of articles from Wikimedia projects[8]. For each page, this project provides the page title, the number of requests (on hourly basis), the language in which the page is written, and the name of the project. The *WikiStats* data set consists of the weekly views of Wikipedia pages in the period from January 2008 to January 2015. We select Wikipedia pages viewed in 2014, then aggregate their page view counts, and those pages viewed more than 1,000 times. Finally, we created two

6 `$http://www.google.com/trends/topcharts?hl=en#date=2014&geo=$`

7 `http://dumps.wikimedia.org/other/pagecounts-raw/`

8 These projects are: wikibooks, wiktionary, wikinews, wikivoyage, wikiquote, wikisource, wikiversity, and wikipedia

Table 26: Number of unique topics per source.

Topics Source	Count
Google global trends	84
Google *.nl* trends	68
WikiStats global	3,293,749
WikiStats .nl	99,396
Real Queries	1,580,386

datasets: the first contains all Wikipedia pages from all domains (*WikiStats* global), and the second contains only pages written in Dutch language (*WikiStats .nl*).

User Queries

Under conditions of strict confidentiality, the KB made anonymized user logs available, collected between March 2015 and December 2015 from users visiting the public digital newspaper archive on a webservice called Delpher. The collection consists of newspapers articles published in the Netherlands since 1618. The data set made available consists of 10 million OCRed newspaper pages in DIDL XML format[9].

Sources summary

We processed all topics from the sources mentioned with the same pre-processing pipeline, which includes lower casing, stopwords (English and Dutch) removal, and the removal of short terms with a length of less than three characters. The resulting dataset statistics are summarized in Table 26.

5.3 ANALYSIS

Using Anchor Text we investigate the coverage of topics in *Common Crawl* (a *breadth-first* crawl), and KB (a *depth-first* crawl). Since Anchor Text usually describes target pages, we first provide a deep analysis of them with regard to their hosts and top-level domains (*TLDs*). Then, we present a detailed analysis of Anchor Text associated with hyperlinks. Finally, we investigate the Anchor Text coverage of topics from the three sources described in Section 5.2.4.

9 http://www.xml.com/pub/a/2001/05/30/didl.html

Table 27: **Analysis of hosts:** For both the target and source pages, we present the absolute count of unique hosts (*first row*), the fraction of hosts from KB_{links} that were found in the corresponding dataset in column header (*second row*), and present information about target hosts that has been crawled in each link dataset (absolute count and percentage).

KB_{links}	CC_{links}	$CC_{links} \cap NL_{tld}$	$CC_{links} \cap KB_{seeds}$	$CC_{links} \setminus KB_{targets}$
Target Hosts				
442,296 100.0%	30,416,854 71.4%	800,957 38.5%	529,962 24.2%	30,100,936 0%
Source Hosts				
31,829 100.0%	9,715,414 57.8%	120,498 28.5%	2,942 8.5%	8,237,940 42.9%
Crawled Target Hosts				
28,640 6.5%	7,260,773 23.9%	67,854 8.5%	2,363 0.4%	6,183,964 20.5%

5.3.1 *Target Pages*

For all link datasets, the number of unique hosts in the target pages is higher than the number of unique hosts of source pages (see Table 27). In KB_{links}, the number of unique target hosts is $442,296$, which is 14 times higher than the number of source hosts ($31,829$). In CC_{links}, the ratio between the target hosts ($30,416,854$) and the source hosts ($9,715,414$) is lower, here, the number of target hosts is only 3 times higher than the number of source hosts. These numbers of source hosts and target host shows the big difference between the two dataset. However, subsets from *Common Crawl* dataset have comparable numbers. The crawling strategy clearly affects the percentage of target hosts that have been crawled. The percentage of the crawled target hosts differ between the link datasets, (see Table 27). For example, only 6.5% of KB_{links} target hosts were crawled, whereas 23.9% of target hosts in CC_{links} were crawled. However, both crawling strategies showed that large fractions of target hosts were not crawled, and we cannot find their raw content. This suggests that the use of target hosts, and Anchor Text as a means to describe them is a valuable resource.

We also looked into the overlap of target hosts between the datasets. A high percentage (71.4%) of target hosts in KB_{links} were also targets of links in CC_{links}. The percentage of overlap decreases to 38.5% after subsetting the *Common Crawl* dataset based on source pages from the *.nl* domain ($CC_{links} \cap NL_{tld}$), and decreases to 24.2% after projecting the KB seeds on CC_{links} ($CC_{links} \cap KB_{seeds}$). Recall, that there is no overlap between KB_{links} and $CC_{links} \setminus KB_{targets}$, be-

Table 28: **TLDs of target pages:** The count of unique TLDs, and the top-10 TLDs.

KB_{links}	CC_{links}	$CC_{links} \cap NL_{tld}$	$CC_{links} \cap KB_{seeds}$	$CC_{links} \setminus KB_{targets}$
nl	com	nl	com	com
com	org	com	org	org
org	net	org	nl	net
net	de	net	net	de
de	info	de	de	info
be	nl	be	be	nl
eu	ru	eu	it	it
info	it	info	ro	ru
fr	fr	it	fr	fr
it	pl	fr	info	pl

cause all links whose target hosts are the same as the target hosts in KB_{links} were dropped from CC_{links}. In terms of the source hosts not only the number of hosts is lower compared to the number of target hosts, but also the overlap between KB_{links} and the other datasets is smaller (see Table 27).

Top-level Domains

Another way of looking at the difference between the link datasets is based on the TLDs of the target pages. The TLDs represent the target domains of the crawled pages. In CC_{links}, a high percentage of links points to the pages from the *.nl* domain, and the majority (60.5%) of the target pages are from the *.com* TLD, see Table 28. The majority of target pages (45.6%) in KB_{links} are from the *.nl* domain, which is expected because the KB crawl was harvested based on websites mainly from the Dutch Web. The target pages in $CC_{links} \cap NL_{tld}$ has the same distribution of top-ranked TLDs of target pages in KB_{links}. In the distribution of TLDs for $CC_{links} \cap KB_{seeds}$, the *.com* is the most prevalent TLD; 49% of target pages belong to this domain, not all websites in the KB seeds were found in *Common Crawl* dataset, only 43.6% (unique) were found. The KB seeds are not all from the *.nl* domain, only 88% of the seeds belong to the *.nl* domain. The remaining seeds (12%) belong to different TLDs: 5% from the *.org* domain, 3.4% from the *.com* domain, 1.2% from the *.net* domain, 0.6 from the *.eu* domain, and 0.5% from the *.info* domain. The distribution of the top TLDs is similar in CC_{links} and $CC_{links} \setminus KB_{targets}$. The only difference is the number of target pages per TLD, which decreases for some TLDs in $CC_{links} \setminus KB_{targets}$ compared to CC_{links}. This is caused by dropping links whose target hosts are the same as the target hosts in KB_{links}. Thus the highest relative decrease was for the *.nl* domain.

Table 29: **Anchor Text summary:** For each link dataset, we present the number of unique Anchor Text, and the overlap of Anchor Text between KB_{links} and the corresponding dataset. Considering all Anchor Text in KB_{links} (%overlap_all), and by considering Anchor Text used at least twice in KB_{links} (%overlap_GT1).

Links Dataset	Count	%overlap_all	%overlap_GT1
KB_{links}	1,581,013	100.0	13.0
CC_{links}	83,920,299	23.6	49.9
$CC_{links} \cap NL_{tld}$	2,613,774	13.7	40.5
$CC_{links} \cap KB_{seeds}$	1,289,803	9.2	26.7
$CC_{links} \setminus KB_{targets}$	61,153,447	15.3	34.4

5.3.2 *Anchor Text*

Some Anchor Text are used by multiple links and the frequency of the Anchor Text represents its popularity in the archive. We processed the Anchor Text with the same pre-processing pipeline we used for the topics (Section 5.2.4) and computed the frequencies of all unique Anchor Text for each link dataset. The number of unique Anchor Text varies strongly among the datasets (see Table 29). When we compared the percentage of overlap between Anchor Text in KB_{links} and all other link datasets based on exact string matching, we found that 23.6% of the unique Anchor Text in KB_{links} exist in the unique Anchor Text of CC_{links}. The frequency of Anchor Text in KB_{links} shows a long tail distribution (Figure 8). A high percentage (87%) of the Anchor Text in KB_{links} occurs only once. We investigated the overlap considering only Anchor Text with a frequency larger than one. This results in an increase of the percentage of overlap between KB_{links} with all datasets. We can use the frequency as threshold to focus on most popular Anchor Text.

5.3.3 *Topic Coverage*

An Anchor Text describes the target page with a brief text which is known to resemble user queries. Therefore analyzing the Anchor Text' overlap with queries is a good proxy for assessing whether the crawls are likely to contain answers to user queries and popular topics. Not all target pages that are linked to from the crawled pages are harvested by the crawler. As mentioned earlier, Web archives are incomplete, and the advantage of Anchor Text is its availability for both crawled and not crawled target pages. In order to investigate the topic coverage, we used exact string matching between pre-processed Anchor Text from the five link datasets with topics from the sources (described in Section 5.2.4). Topic coverage varies among

Table 30: **Topic Coverage:** for each link dataset, we present the absolute count and the fraction (%) of found topics in each topic source, where the fraction is the number of matched topics to the total number of topics in the corresponding source. The %*lost* under $CC_{links} \setminus KB_{targets}$ is the relative not found topics, these topics were found in CC_{links} but in $CC_{links} \setminus KB_{targets}$.

Topics Source	KB_{links}		CC_{links}		$CC_{links} \cap NL_{tld}$		$CC_{links} \cap KB_{seeds}$		$CC_{links} \setminus KB_{targets}$		
	count	%	count	%	count	%	count	%	count	%	%lost
Google global trends	24	28.6	51	60.7	25	29.8	23	27.4	51	60.7	0.0
Google *.nl* trends	22	32.4	27	39.7	25	36.8	18	26.5	24	35.3	-11.1
WikiStats global	80,043	2.4	1,376,222	41.8	122,659	3.7	116,259	3.5	1,122,767	34.1	-18.4
WikiStats .nl	24,726	24.9	48,825	49.1	31,742	31.9	19,098	19.2	43,304	43.6	-11.3
Real Queries	26,099	1.7	77,152	4.9	38,033	2.4	15,839	1.0	66,874	4.2	-13.3

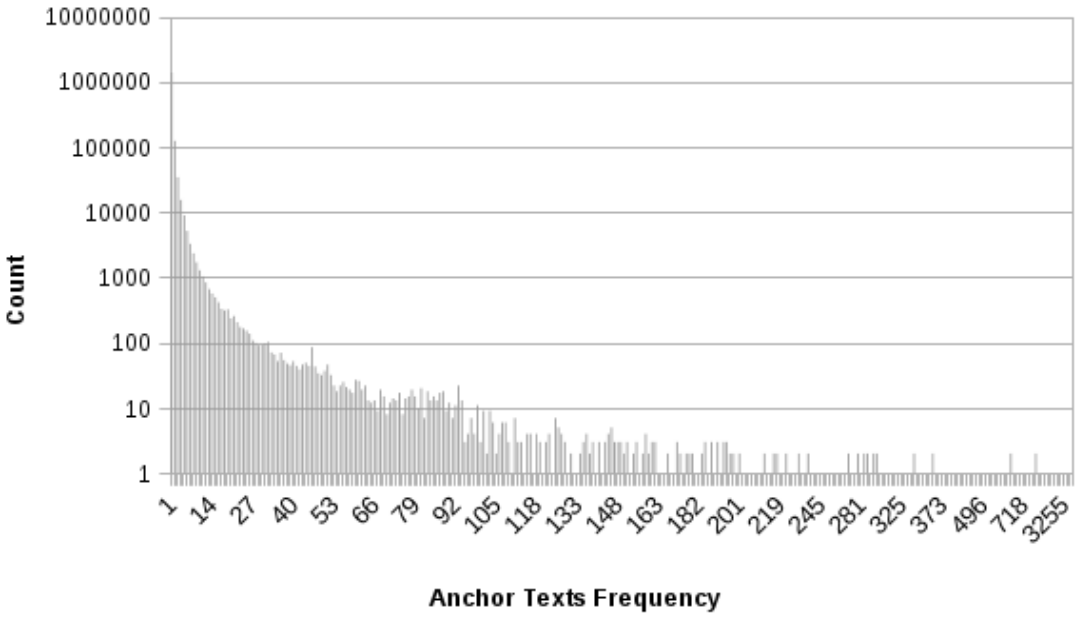


Figure 8: Anchor Text frequency distribution of KB_{links} in log scale representation.

the datasets for the different sources of topics, (see Table 30). For some cases we found high coverage, for example Anchor Text from CC_{links} matched 60.7% of Google global trends and 49.1% of the Dutch Wikipedia pages in the *WikiStats*. After sorting all Anchor Text in descending order based on their frequencies, we investigated the relation between the percentage of topics covered and the frequency (popularity) of Anchor Text. We report on exact string matches between the top Anchor Text and both, Wikipedia titles and real user queries, considering different rank cutoffs c; c = 1k, 10k, and 100k. The percentage of matched Anchor Text with *WikiStats* (global and *.nl* domain) decreases as c increases for all link datasets (see Table 32). The lowest overlap corresponds to the case when all Anchor Text are used to match Wikipedia titles.

In general, the percentage of overlap between Anchor Text from the different datasets and the user queries is low. For example, we found that only 1.7% of the user queries had a match in KB_{links} when we applied exact string matching with all Anchor Text. We found the highest percentage of overlap with user queries (4.9%) for Anchor Text in CC_{links} (see Table 30). When we compared the top-c Anchor Text instead of the complete set of Anchor Text, we found a relation between the top ranked Anchor Text and the percentage of the topic coverage. A high percentage of the most frequently used Anchor Text matched user queries, and the percentage of overlap decreases while the cutoff c increases, see Table 33. We found the highest overlap of topics and Anchor Text in CC_{links}, suggesting that the *breadth-first* crawl covers more topics than the *depth-first* crawl. This result holds for both, the

Table 31: **Unique Topic Coverage in KB_{links}:** in comparison with topics found in other datasets. Under every link dataset x, we present the percentage of topics found in the KB but not found in x (*first column*), and the percentage of topics found in x but not found in KB_{links} (*second column*).

Topics Source	CC_{links}		$CC_{links} \cap NL_{tld}$		$CC_{links} \cap KB_{seeds}$		$CC_{links} \setminus KB_{targets}$	
Google global trends	0.0%	54.9%	29.2%	32.0%	33.3%	30.4%	0.0%	54.9%
Google *.nl* trends	0.0%	18.5%	9.1%	20.0%	31.8%	16.7%	13.6%	20.8%
WikiStats global	6.1%	94.5%	46.1%	64.8%	56.3%	69.9%	12.2%	93.7%
WikiStats .nl	16.5%	57.7%	28.2%	44.0%	53.6%	39.9%	26.4%	58.0%
Real Queries	22.3%	73.7%	31.4%	53.0%	60.4%	34.7%	31.8%	73.4%

Table 32: **The fraction of top-c ranked Anchor Text matching document titles in *WikiStats***, fraction in percentage (*num. matches/c*). In addition to the percentage of matching when all Anchor Text are used (*num. matches/num. all Anchor Text*).

Links Dataset	*WikiStats* global			
	top-1k	top-10k	top-100k	all
KB_{links}	44.6	39.1	22.1	5.1
CC_{links}	51.3	56.9	38.2	1.6
$CC_{links} \cap NL_{tld}$	45.2	37.3	24.7	4.7
$CC_{links} \cap KB_{seeds}$	70.6	65.6	33.9	9.0
$CC_{links} \setminus KB_{targets}$	71.2	63.5	28.1	1.8
Links Dataset	*WikiStats .nl* domain			
	top-1k	top-10k	top-100k	all
KB_{links}	32.4	24.3	10.2	1.6
CC_{links}	11.5	14.4	7.4	0.1
$CC_{links} \cap NL_{tld}$	32.0	23.9	11.8	1.2
$CC_{links} \cap KB_{seeds}$	32.0	23.3	8.3	1.5
$CC_{links} \setminus KB_{targets}$	19.5	14.7	4.2	0.1

global and the national (*.nl*) topics. Focusing on the Dutch part of the *Common Crawl* dataset ($CC_{links} \cap NL_{tld}$) showed that this part covers more topics than topics covered in KB_{links}. However, the comparison is based on the absolute count of found topics in each links dataset. That does not necessarily mean that all topics covered by KB_{links}, are identical with those found, for instance, in CC_{links}. For all topic sources, we analyzed the topics that were found in KB_{links} but not in the other datasets (see Table 31). For example, we found that all Google trends (both the global and the *.nl* domain) that were found in KB_{links}, were also found in CC_{links}. On the other hand, 54.9%

Table 33: **The fraction of top-c ranked Anchor Text matching user queries**, same notation as in Table 32.

Links Dataset	Real Queries			
	top-1k	top-10k	top-100k	all
KB_{links}	26.4	23.2	9.7	1.7
CC_{links}	17.2	18.2	7.0	0.9
$CC_{links} \cap NL_{tld}$	33.5	26.0	12.8	1.5
$CC_{links} \cap KB_{seeds}$	30.0	20.5	6.9	1.2
$CC_{links} \setminus KB_{targets}$	28.5	23.0	6.9	0.1

of Google's global trends and 18.4% of Google's *.nl* trends found in CC_{links} were not found in KB_{links}. Regarding the *WikiStats* dataset, not all topics found in KB_{links} were found in CC_{links}. The percentage of topics that were found in KB_{links} is higher for the Wikipedia pages from the *.nl* domain (16.5%), while 6.1% of Wikipedia pages (global) found in KB_{links} were also found in CC_{links}.

These results suggest that Anchor Text can be used as a resource for finding topics that were popular with users from the past. The coverage of topics was higher for the most frequently used Anchor Text in the crawls. Anchor Text from the *breadth-first* crawl covers more topics than the Anchor Text from the *depth-first* crawl. However, some topics were only covered by the *depth-first* crawl.

5.4 CONCLUSIONS

We studied the influence of the crawling strategy on the coverage of topics that were of interest to users on the Web. We performed our analysis on two Web crawls created by following different crawling strategies; the *Common Crawl* dataset, (a *breadth-first* crawl) collected from the entire Web, and the KB dataset (a *depth-first* crawl) harvested by the KB based on manually selected websites). We made use of Anchor Text to investigate the topic coverage in the two crawls. We extracted Anchor Text from the raw content of documents in crawls, and compared them with other sources that identify popular topics on Web at the time of the crawls (2014). The two crawls differ in terms of scope. While *Common Crawl* covers domains from the entire Web, KB covers mainly the Dutch domain. Therefore, we used different sources as a proxy of topics that were popular in 2014, both worldwide (entire Web) and national (*.nl* domain).

Using exact string matching between Anchor Text and topics from different sources, we found that the percentages of matches vary between the topic sources and the two crawls. For example, CC_{links} covers 61% of Google global trends, and 5% of real queries (submitted by users to the search system of the Dutch digital newspaper archive). KB_{links} covers 32% of Google *.nl* trends, and 2% of the real queries. This suggests that Anchor Text are a useful resource for investigating popular topics from the past. We found a correlation between the frequency of Anchor Text in the archive and the percentage of topic matches.

When we compared the topic coverage between the *Common Crawl* and the KB datasets, we found that the percentage of overlapping topics is higher in the *Common Crawl* dataset, for both global and *.nl* topics. This result holds for the $CC_{links} \cap NL_{tld}$ (only focusing on links in *Common Crawl* originating from the *.nl* domain). More over, using the $CC_{links} \cap KB_{seeds}$ (was created using KB seeds to subset CC_{links}) has comparable result to KB_{links}. However, not all topics

found by the *depth-first* crawl were found by the *breadth-first* crawl. We conclude that the coverage in the *breadth-first* crawl is higher even for topics of national interest, but there are topics that are covered only by *depth-first* crawl.

In future work, we can investigate the topic coverage in the crawls taking the importance of topics into account, in this analysis all topics were weighted equally.

6

QUANTIFYING RETRIEVAL BIAS IN WEB ARCHIVE SEARCH

A Web archive usually contains multiple versions of documents crawled from the Web at different points in time. One possible way for users to access a Web archive is through full-text search systems. However, previous studies have shown that these systems can induce a bias, known as the *retrievability bias*, on the accessibility of documents in community-collected collections (such as TREC collections). This bias can be measured by analyzing the distribution of the *retrievability scores* for each document in a collection, quantifying the likelihood of a document's retrieval.

We investigate the suitability of retrievability scores in retrieval systems that consider every version of a document in a Web archive as an independent document. We show that the retrievability of documents can vary for different versions of the same document, and that retrieval systems induce biases to different extents.

We quantify this bias for a retrieval system which is adapted to handle multiple versions of the same document. The retrieval system indexes each version of a document independently and we refine the search results using two techniques to aggregate similar versions. The first approach is to collapse similar versions of a document based on content-similarity. The second approach is to collapse all versions of the same document based on their URLs. In both cases, we found that the degree of bias is related to the aggregation level of versions of the same document.

Finally, we study the effect of bias across time using the retrievability measure. Specifically, we investigate whether the number of documents crawled in a particular year correlates with the number of documents in the search results from that year. Assuming queries are not inherently temporal in nature, the analysis is based on the timestamps of documents in the search results returned using the retrieval model for all queries. The results show a relation between the number of documents per year and the number of documents retrieved by the retrieval system from that year. We further investigated the relation between the queries' timestamps and the documents' timestamps. First, we split the queries into different time-frames using a one-year granularity. Then, we issued the queries against the retrieval system. The results show that temporal queries indeed retrieve more documents from the assumed time-frame. Thus, the documents from the same time-frame were preferred by the retrieval system over documents from other time-frames.

6.1 INTRODUCTION

Indexing and retrieving documents from a Web archive can be challenging. Web archive collections are different from conventional static Web collections. The main reasons are the continuously increasing size of Web archives and the existence of multiple versions of the same document collected at different moments in time. The different versions may appear multiple times in search results and thereby render other documents inaccessible for a user. Despite these challenges, Web archive initiatives make an effort to make their collections better accessible. For example, Gomes et al. conducted a survey in 2010 on 42 Web archive initiatives around the world (26 countries) [96]. They found that 89% of the initiatives support access to the Web archive by a given URL, 79% support searching meta-data, and 67% provide *full-text* search over their archives. The same survey was conducted again in 2014 in order to observe the change in Web archiving since 2010 [72]. They noticed an increase in the number of initiatives (68) and the number of countries involved in Web archiving (33 countries). However, in terms of access methods, the results of 2014 are the same as those for 2010.

Previous studies showed that applying existing Information Retrieval (IR) models on Web archives leads to unsatisfactory results [70, 69]. Measuring the effectiveness of IR systems can be done using test collections. A test collection consists of a set of topics (queries), a document collection, and a set of relevance assessments. Costa and Silva extended this approach by taking the characteristics of Web archives into account [69]. Their approach includes the design of a test collection and constructing topics from the users' query log of a functioning Web archive search system. Getting relevance judgments, however, is a costly process. An additional complication is the dependency on query logs, as they are seldomly available for research.

To complement standard methods of IR evaluation, that focus on the assessment of efficiency and effectiveness of IR systems, Azzopardi et al. introduced *retrievability* as a measure for potential bias in the access of documents in a collection [39]. The retrievability score of a document counts how often the document is retrieved when a large representative set of queries is issued on the retrieval system. The overall bias in the scores among all documents in the collection induced by a retrieval system can be quantified using measures such as the Lorenz Curve [91] and the Gini Coefficient [91]. While the Lorenz Curve can be used to visualize the bias, the Gini Coefficient can be used to quantify the extent of bias for different experimental conditions.

We follow an approach similar to [39] to study how retrievability can be used to quantify retrieval bias induced by different retrieval

systems on a subset of the Dutch Web archive collection from the National Library of The Netherlands[1] (KB).

Our main goal is to investigate how to use retrievability to evaluate a Web archive retrieval system, and how the number of document versions and the method of aggregation of crawls influence the retrieval bias in the Web archive.

Specifically, we address the following research questions:

RQ4 *What can we learn about Web archive access from studying the collection using a measure of retrievability?*

RQ4.1 *Is access to the Web archive collection influenced by a retrievability bias? Can we evaluate and compare retrieval systems on the Web archive collection using the retrievability measure to quantify their retrieval bias?*

We follow the approach of [39] to quantify the overall bias imposed by different retrieval systems using the Gini Coefficient and the Lorenz Curve constructed using retrievability scores of documents in the collection.

RQ4.2 *How does the number of versions of documents in the Web archive collection influence the retrievability bias of a retrieval system?*

The number of versions per document in the archive varies, for example because documents have been crawled with different frequency or because they were added to the crawler's seed list at different points in time. We show how the multiple versions impact the retrieval bias when the granularity of retrieval in the search results is the document's version. We compute the retrievability score of a document by accumulating the retrievability score of its versions; a document with more versions gets higher retrievability score. Then, we show the change in bias when the multiple versions are handled by the retrieval system using two approaches to collapse documents' versions: first, based on their content-similarity; second, based on their URLs.

RQ4.3 *Does a retrieval system favor specific subsets of the collection?*

The Web archive collection of the KB consists of snapshots of websites from different points in time spanning four years. Therefore, we investigate what subset of the archive is most affected by retrieval bias.

1 www.kb.nl

The remainder of the chapter is organized as follows. After discussing related work (Section 6.2) we describe our approach to answer the research questions introduced in this section (Section 6.3). We discuss the experimental setup in detail in (Section 6.4) and answer research questions *RQ*4.1-4.3 in Sections 6.5, 6.6, and 6.7, respectively. Finally, we discuss conclusions drawn from our findings (Section 6.8).

6.2 RELATED WORK

Understanding the information needs of Web archive users is an important step towards developing good access methods for Web archives. Several studies showed that *full-text* search is preferred [67, 68, 96, 141]. This shift from single URL search to search interfaces was described as a turning point in the history of Web archives [50].

Research in *temporal IR* aims to exploit temporal information in documents and queries for better query understanding and time-based ranking [35, 62, 112]. In [69], Costa and Silva created a temporal test collection from the Portuguese Web Archive [95], to enable evaluation of temporal methods in IR. A test collection consists of queries (topics), documents, and the judgments by users of their relevance to the queries. When a new system is built then its effectiveness can be measured based on the test collection using evaluation metrics such as precision (for example *P@10*). The collection developed by Costa and Silva consists of crawls in the period from 1996 to 2009. The queries (topics) were selected from query logs, and the documents retrieved by the retrieval system were manually judged. Their method extends the Cranfield paradigm with consideration of the temporal aspect of Web archive collections. Other studies used crowdsourcing to collect relevance judgments. For example, Berberich et al. used Amazon Mechanical Turk to collect queries and relevance assessments [51].

Retrievability was introduced to measure how likely a document is to be retrieved given an IR system [40, 41, 39]. Computing the retrievability scores requires the availability of a large query set, but without the need for relevance judgments. Queries can be simulated by drawing them from the content of documents in the collection. The retrievability score of a document $r(d)$ gives an indication of how retrievable the document is compared to other documents in the collection. It is computed by accumulating the number of times this document appears in the ranked list provided for all queries, at a given cutoff rank. In order to quantify the retrievability bias across all documents in the collection, the Lorenz Curve [91] is used to visualize the bias and the Gini Coefficient [91] is used to summarize the bias. In economics, the Lorenz Curve is used to visualize the distribution of wealth or income of a population. If the wealth or income is

equally distributed in the population, the accumulative distribution is a diagonal line (called the line of equality). The larger the inequality is within a population, the more the curve deviates from the equality line. The Gini Coefficient summarizes the overall inequality into a value which ranges from zero (perfect equality) to one (perfect inequality). The Gini Coefficient quantifies the retrievability inequality among documents. In the context of retrievability, the population corresponds to the document collection and wealth corresponds to the retrievability scores.

Retrievability has been used to compare different retrieval models based on the bias they impose on a given collection, and to study whether the retrieval system favors documents with particular features. For example, the system might favor long documents over shorter documents. In the following, we discuss a few studies that used retrievability. Retrievability was applied in the patent search domain [43, 47], which is recall-oriented, to quantify the retrieval bias of retrieval systems on the patent collection. The correlation between retrievability and the query set was considered in several studies. Based on a limited set of queries, the correlation between retrievability score and query relevance to the document [2] was analyzed [45]. Their experimental results showed that 90% of highly retrievable documents when all queries were considered, are not highly retrievable considering only their relevant queries. The influence of query characteristics on retrieval bias was explored in [48]. They showed that different query characteristics increase or decrease the retrieval bias differently. Query expansion was used to improve document's retrievability [46].

Other studies investigated the relation between a system's retrieval bias and its effectiveness. For example, Azzopardi et al. [37] showed that a positive relation exists between effectiveness and retrievability. Measuring effectiveness using precision at 10 (P@10) & Mean Average Precision (MAP), the results showed that as the effectiveness increases, the retrievability bias tends to decrease. This relationship between retrievability and effectiveness has been used to tune systems [158]. Bashir and Rauber investigated the impact of query expansion on the retrievability bias [46]. They showed that standard query expansion methods caused an increase in effectiveness and retrieval bias. They explained the increase in retrieval bias due to the assumption of query expansion methods that the top-ranked documents are relevant. However, some documents in the top-ranked results might be noise. Therefore, in order to decrease the retrieval bias, they proposed a query expansion approach based on document clustering, and they showed that their approach reduces the bias.

2 The relevance of the document to each query in a small sample was assessed by experts

6.3 APPROACH

We explore how we can use retrievability to assess the retrieval bias of retrieval systems providing access to four years of the Dutch Web archive. In order to investigate our first research question, *RQ4.1*, we use three well-known IR models and two large query sets. For every model and query set, we compute the retrievability score ($r(d)$) for document versions at different rank cutoffs c. Parameter c represents the willingness of the user to explore a certain number of documents in the search results, therefore it is independent from the retrieval model. In our study we experiment with $c = 10$, 20, 30, 40, 50, 100, and $1,000$. Users are known to rarely evaluate more than the first 10 search results, however, we also consider high values for c to find out whether the inequality bias would still exist if the users were willing to explore higher numbers of results. In order to allow the comparison of the retrieval models in terms of retrieval bias they impose on the documents, we need a measure to quantify the overall bias given a collection, a query set, and a retrieval system. We use the Gini Coefficient to summarize the retrieval bias, and the Lorenz Curve to visualize the retrieval bias, following [39].

A certain fraction of documents is *not-retrieved* by any of the retrieval models. This fraction is especially high for smaller c's, and has a strong influence on the overall bias measured by the Gini Coefficient. Therefore, we compute two variants of the Gini Coefficient. In the first variant, all documents in the collection are included; if a document is not retrieved by the model, its retrievability score is zero ($r(d) = 0$). Here, the number of documents is the same for all models at all c's (number of retrieved documents plus number of not-retrieved documents = whole collection). In the second variant, only documents that are retrieved using at least one of the three retrieval models at a given c are considered. We do this by creating a union set of unique documents retrieved using at least one of the three models at the given c (*3Models_union_c*) for each query set. If a document was retrieved using model A, but not with model B, then the retrievability score of that document given model B is assigned a value of 0 ($r_B(d) = 0$). The number of documents will be the same for all models at the same c (num. retrieved plus num. not-retrieved = *3Models_union_c*). Therefore, this can still be considered to provide a fair comparison across the retrieval models for a given c. Using the second variant will reduce the impact of a high fraction of documents with $r(d) = 0$. A model that does not retrieve a large number of documents that were retrieved using other models, will get a higher Gini Coefficient, that is, it is considered to be more biased.

In order to understand the relation between the retrievability scores and the ability to find a document in the collection, we use a known-item-search setup based on the approach proposed in [38, 42].

We quantify the impact of multiple versions of the same document on the inequality of retrieval bias, *RQ4.2*. First, we investigate the retrieval of all versions of a document. At indexing and retrieval time we consider the document's version as an independent document. In order to check how that affects the document's retrievability, we compute the retrievability of a document by aggregating the retrievability scores of its versions retrieved at a given c, and thus the overall bias imposed by the model. Second, we collapse similar versions of the same document and again compute the retrievability score and the overall bias. Third, to explore the impact of the number of versions on the bias, we linearly combine the scores given by the models with a prior based on the number of versions. This allows us to measure retrieval bias at the granularity of the document, instead of a specific version.

Finally, we address our last research question, *RQ4.3*. Our Web archive collection is an accumulation of several crawls over time. We are interested in whether the bias imposed by a given retrieval system, among subsets based on the time of crawling, correlates with the number of crawled documents in that year. To explore this research question, we focus on the documents retrieved using the *BM25* model; as we show in the results, it provides the least bias. Using the timestamps of the crawling time associated with documents, we split the search results for *BM25* into four subsets at different c's, and then measure the retrieval bias per subset.

6.4 EXPERIMENTAL SETUP

In Section 6.4.1, we describe the components used to measure retrievability on the Web archive collection. In Section 6.4.2, we describe the known-item search setting to investigate the relation between retrievability score of a document and the difficulty level of finding that document.

6.4.1 *Retrievability Experimental Setup*

First, we introduce the Dutch Web archive collection (Section 6.4.1.1). Then, we describe how we pre-processed and indexed the collection (Section 6.4.1.2). After that, we discuss how we designed the query sets that are used to retrieve documents from the collection (Section 6.4.1.3). Finally, we discuss how to measure retrievability scores and how to quantify the overall bias imposed by a given retrieval model (Section 6.4.1.4).

Table 34: Summary of the archived objects over the years, with more details on documents of *text/HTML* content-type. The mean value of number of versions was computed by dividing the total number of document versions crawled per year over the unique number of documents (URLs). The number of original URLs for *All* years is the number of unique URLs in the four years.

Year	archived objects all types	text/HTML documents			
		all versions (mementos)	%	original URLs	Mean (#versions)
2009	17,014,067	12,232,831	71.9	9,764,370	1.25
2010	38,157,308	22,596,291	59.2	17,093,870	1.32
2011	53,604,464	30,275,150	56.5	19,491,258	1.55
2012	38,865,673	19,464,431	50.1	13,191,771	1.48
All	147,641,512	84,568,703	57.3	47,836,163	1.77

6.4.1.1 *Data Set*

In their Web archive, the KB preserves a growing seed set of currently more than 10,000 websites [140]. For our research, the KB provided us with a subset of the Dutch Web archive that has been harvested between February 2009 and December 2012, consisting of 76,828 Archive (*ARC*[3]) files. Each ARC file contains multiple archived records (content plus the response header), which yields a total of 148M documents. Table 34 shows the total number of archived objects, raw count and the percentage of *text/html*. We refer to *text/HTML* content-type objects as documents. These documents form our collection D on which we focus our analysis. Every crawled document has its own URL and the timestamp of the crawling time in addition to its content on the Web at the time of the crawl. Every document d may have multiple versions crawled at different points in time t_i,

$$d := \{d_v^{t_1}, d_v^{t_2}, ..., d_v^{t_n}\}$$

where $d_v^{t_1}$ is the *document's version* crawled at time t_1. The mean value of number of versions (total number of versions over the number of unique documents based on URLs) increases over the years, as more crawls have been added to the archive (see Table 34). The distribution of the number of versions per document is skewed (see Figure 9, in a log scale).

6.4.1.2 *Pre-Processing & Indexing*

Pre-processing consists of removing HTML tags, tokenization, removing stopwords, removing terms of length less than 3 characters, re-

3 http://archive.org/web/researcher/ArcFileFormat.php

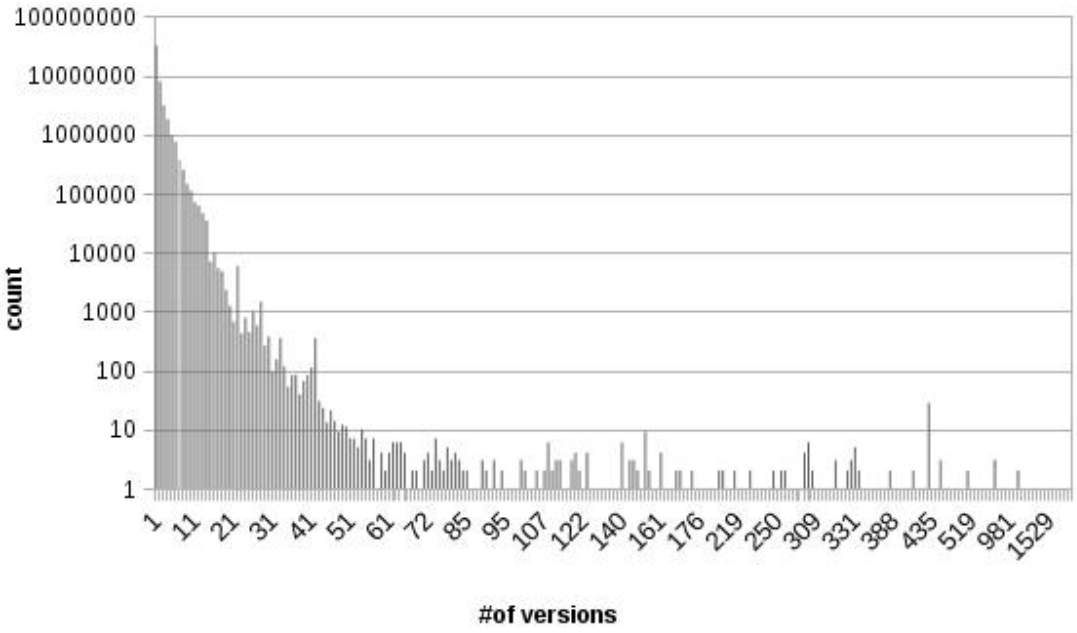


Figure 9: Distribution of number of versions of documents in the Dutch Web archive collection in log scale representation.

moving numbers with fewer than 4 digits, and stemming. For every document's version $d_v^{t_i}$, we keep the following data:

$$d_v^{t_i} := \{\text{URL, docId, crawl-date, pre-processed-content}\}$$

where the docId is a unique identifier defining the document's version, while the URL is the same for all versions of the same document. We used the Lemur toolkit[4] to index our collection. The documents in our collection are in Dutch, but unfortunately, a Dutch stemmer is not available in the Lemur toolkit. Therefore, we applied stemming in the pre-processing stage[5] and switch off stopword removal and stemming at indexing time (as these have already been applied in the pre-processing stage). The index granularity is the document's version $d_v^{t_i}$. For indexing and retrieval, we used the same IR systems as [39], motivated by their widespread application in IR [128]: *BM25*, *TF*IDF* and *LM1000* (Language Modeling with Bayes Smoothing, $\mu = 1,000$).

6.4.1.3 *Query Set*

In order to compute the retrievability score of all documents in the collection, we need a set of queries to run against a given retrieval system. Ideally, we would use queries collected from users searching the collection. Unfortunately, such a query log is not available

4 http://www.lemurproject.org/

5 `https://lucene.apache.org/core/4_3_0/analyzers-common/org/apache/lucene/analysis/nl/DutchAnalyzer.html`

for the Web archive. However, there are reasonable alternatives for generating the query set. First, we follow the approach used in [39] by simulating the queries from the content of the documents in the collection. Second, we use the hyperlink's Anchor Text in the Web archive. One of the defining properties of the Internet is its hyperlink-based structure. The structure of the Web graph is defined by its hyperlinks which consist of a source URL, a destination URL, and an Anchor Text describing the destination. The hyperlink structure is a rich source of information about the content of a Web collection which has been widely used, especially in the context of Web retrieval, including the *PageRank* algorithm for ranking Web documents [139], and Kleinberg's approach to infer hubs and authorities [117]. Anchor Text has been widely used in the IR field to improve search effectiveness [73, 85, 90, 110, 119, 121, 132]. Empirical studies have shown that Anchor Text exhibits characteristics similar to both user queries and document titles [86]. Language models generated from document titles also can be used as an approximation of a user query language model [108]. In addition to this, Anchor Text is available not only for pages in the archive, but also for pages that have not been archived when there are pointers to them from pages in the Web archive [115, 106, 142].

simulated query sets The first choice for generating a large set of queries is to draw them from the text content of documents in the collection following [39]. Their approach exploits the idea behind query based sampling [61], a method that summarizes the content of a database in a non-cooperative distributed search setting starting with a set of keywords. From the pre-processed documents, as described in Section 6.4.1.2, we generate queries of one or two terms. The single-term query set was constructed by taking the most frequent 2 million terms in the collection. The frequencies of the single-term queries range from 5 to 204,517,438. The bi-term query set was constructed by generating all possible two consecutively occurring terms (*bigrams*) from the content of the pre-processed documents. Then, we selected the first 2 million bigrams after ranking them based on number of occurrences. The frequencies of the bi-term queries range from 20 to 35,490,632. The single-term and bi-term queries constitute query set Q_s (4 million queries).

anchor text query set The second set of queries consists of Anchor Text constructed from links which we extract from the collection. A link consists of the source URL (the URL of the page where the link was placed), target URL (the URL of the page that the link points to), and the Anchor Text of the link (a short text describing the target page). To extract the links from the archive, we process all archived web objects contained in the archive's ARC files. During the process-

Table 35: Summary of the query sets.

Query Set	# of queries	Mean query length	#of terms
Q_s	4,000,000	1.5	2,000,000
Q_a	1,763,668	2.4	755,589

ing, JSoup[6] was used to extract links. For each found anchor link, we keep the source URL, the target URL, and the Anchor Text. We extract the crawl date from a document's metadata, and combine the date with the link information. More precisely, we keep:

<sourceURL, targetURL, anchorText, crawlDate>

We only use the Anchor Text from external links, where the domain-name of the source URL is different from that of the target URL (an inter-domain link). Different seeds are harvested at different frequencies: while most sites are harvested only once a year, some sites are crawled more frequently. Therefore, we deduplicate the links based on their values for source, target, Anchor Text, and the year of the crawl date. We aggregate the link entries by Anchor Text and sort them based on their frequency (number of times used to point to the target). Finally, we apply stopword removal and stemming; we refer to this query set as Q_a.

summary of query sets Table 35 provides the total number of queries, average query length based on the number of terms used per query, and the total number of terms used in each query set (vocabulary of each query set). The number of terms in the vocabulary of the Q_s query set is high. Recall that the simulated queries were extracted from the content of the documents after pre-processing. The terms that were excluded are the Dutch stopwords, terms of length less than 3 characters, and numbers of less than 4 digits. Terms that pass these filters are included, such as numbers, for example dates, telephone numbers, and terms in different languages. After calculating the frequency of terms in the Q_s query set (i.e., the number of queries using each term), we found that a high percentage (45%) of terms were used by one query.

We found that there are 357,258 terms in the overlap between the vocabulary of the two query sets, which is 47.3% of terms in the Q_a vocabulary, and 18.0% of the Q_s vocabulary. To get insights whether the terms in the overlap are the most or least frequent terms, we sorted the vocabulary terms of each query set in descending order based on their frequency; a term frequency is the number of queries using that term. Then, we computed the percentage of overlap at

6 http://jsoup.org/

Table 36: Percentage of overlap between the vocabulary of the query sets at different cutoff levels after sorting terms in descending order.

top-c	% of overlap
top-10k	62.1
top-50k	57.1
top-100k	49.8
top-200k	43.0
top-300k	34.3
top-500k	27.3

Table 37: Query length distribution of queries in the Q_a query set.

query length	number of queries	percentage
1	397,892	22.6
2	578,819	32.8
3	444,093	25.2
4	247,381	14.0
5	84,463	4.8
6	10,993	0.6

different rank cutoff levels. The percentage of overlap was decreasing by increasing the cutoff of the top frequent terms (see Table 36). In terms of query length, the mean query length (number of terms) of the Q_s query set is 1.5 terms; half of the queries are single-term, and the other half are bi-term queries. The mean query length is 2.4 terms for the Q_a query set. 22.6% of the queries are single-term queries, 32.8% are bi-terms queries, and 25.2% are three-terms queries (see Table 37).

6.4.1.4 *Retrievability Assessment*

For each of the three IR models discussed above, we issue queries in the query set Q, where $\{Q := Q_s, Q := Q_a\}$. For each $q \in Q$, we collect a ranked list of $1,000$ documents. Each document in the ranked list has an associated score representing its estimated relevance to the query, and a number representing its position in the ranked list for the retrieval model. The retrievability $r(d)$ of a document d with respect to an IR model given a query set Q is defined as follows (see also [39]):

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, \{c, g\}) \tag{5}$$

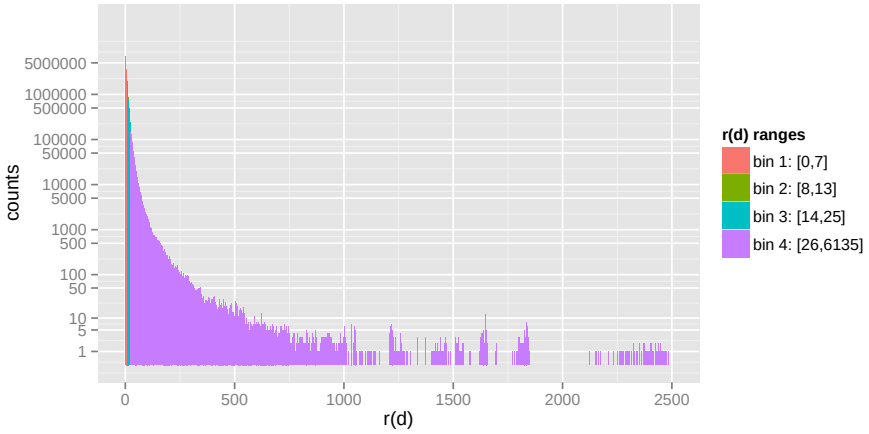
where q is a query from a query set Q, k_{dq} is the rank at which document d is retrieved for q, and $f(k_{dq},\{c,g\})$ is the access function which indicates how retrievable is d for given q at rank cutoff c. The parameter c represents the effort that the user makes to explore more documents from the provided ranked list. In other words, $f(k_{dq},\{c,g\}) = 1$ if d is retrieved for q in the given c, otherwise $f(k_{dq},\{c,g\}) = 0$. For each query set and retrieval model, we compute the retrievability score for all documents in the collection using different $c \in \{10, 20, 30, 40, 50, 100, 1,000\}$. Based on Equation 5, the more queries retrieve d at a given c, the higher is $r(d)$. The o_q coefficient represents the importance of the query. If we have a real user log, then this coefficient can be the likelihood of using the query; this relates to the number of times the query was issued by users. In our analysis, we consider $o_q = 1$ for all queries as the queries were simulated from the collection, not issued by real users.

In order to quantify the global retrievability bias across all documents in the collection, we follow [39] in using the Lorenz Curve [91] and the Gini Coefficient (G) which was proposed to summarize the bias in the Lorenz Curve [91]. If a system imposes no bias on the collection and all documents are equally retrievable, then $G = 0$. On the other extreme, if $G = 1$, then the same document is always retrieved for every $q \in Q$ and the remaining documents in the collection are never retrieved. The Lorenz Curve curve visually shows the retrieval bias variation between the retrieval models. The more the curve of a retrieval model deviates from the linear line of equality, the greater the bias imposed by that retrieval model.

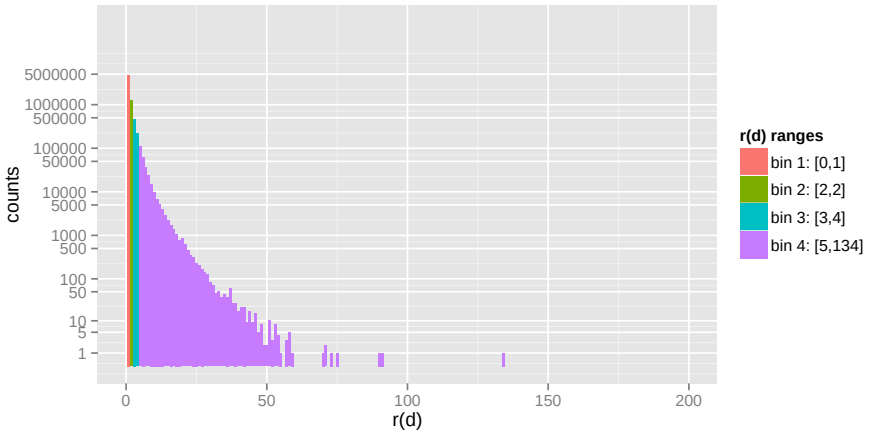
6.4.2 *Known-Item Search Setup Based on Retrievability Scores*

In our known-item search experiment, a query formulated from a document (target document) is used to find that document, and the Mean Reciprocal Rank (MRR) is computed based on the position of the target document. In order to validate the relation between a document's retrievability score and the difficulty level of finding that document, we split documents into bins after sorting them based on their retrievability scores. We perform a known-item search experiment on queries simulated from the results of *BM25*, based on the two query sets, Q_a and Q_s. We select a high c, as more documents were retrieved ($r(d) > 0$); precisely we select $c = 100$, and $c = 1,000$. Based on the Q_s query set, *BM25* retrieved 50.2% of the documents in collection at $c = 100$, and 71.8% at $c = 1,000$. Based on the Q_a query set, 34.7% was retrieved at $c = 100$, and 64.9% was retrieved at $c = 1,000$ by *BM25*. We perform the known-item experiment based on the following steps:

1. Based on the documents' retrievability scores (which will be discussed in Section 6.5), we divide the collection into 4 bins. In ad-



Q_s, $c = 100$



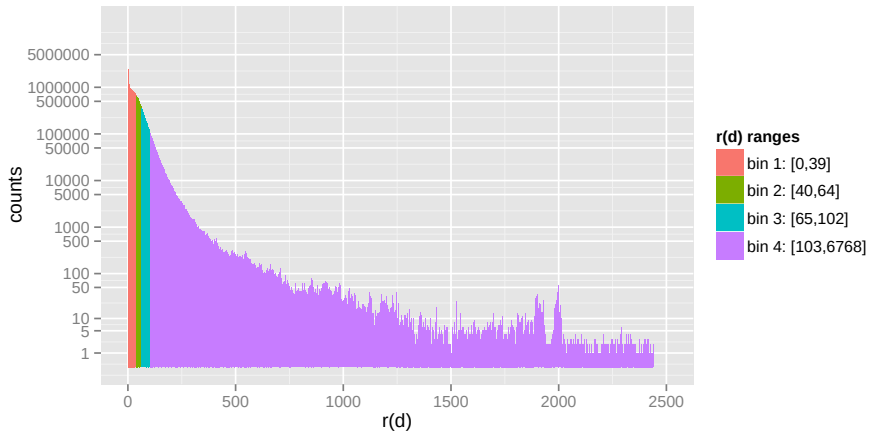
Q_a, $c = 100$

Figure 10: Distribution of retrievability scores $r(d)$ for *BM25* based on all documents in the collection at $c = 100$.

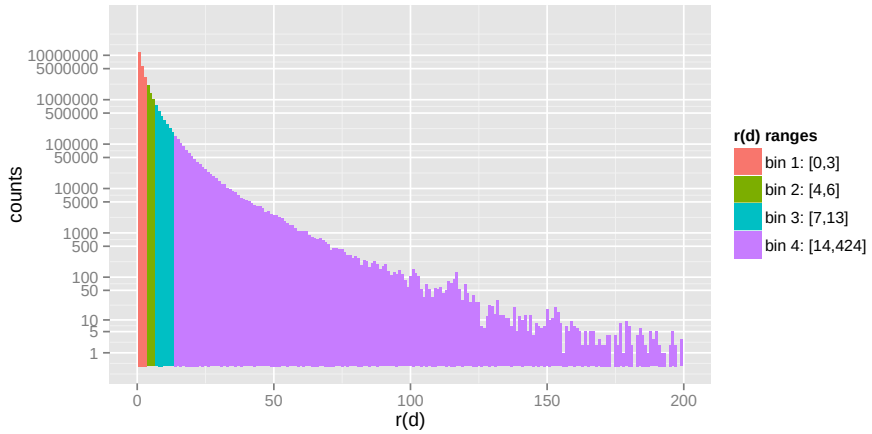
dition to the retrieved documents, we include the non-retrieved ($r(d) = 0$) as they are the most difficult to retrieve.

1.a) Azzopardi et al. sort the documents in ascending order based on their retrievability scores and divide them into 4 bins [39].

1.b) In our setup this way of binning would mean that the non-retrieved documents dominate the first bins. The fraction of non-retrieved documents at $c = 100$ is 49.8% for the Q_s query set, which would mean that two bins would contain only those. The percentage of ($r(d) = 0$) based on the Q_a is higher, 35.1% at ($c = 1,000$) and 65.3% at ($c = 100$). Instead, we chose to partition the documents based on the *wealth* distribution. The wealth is computed by multiplying each retrievability score by the number of documents having that retrievability score. We accumulate the wealth until 25% of the total wealth is reached



Q_s, $c = 1{,}000$



Q_a, $c = 1{,}000$

Figure 11: Distribution of retrievability scores $r(d)$ for *BM25* based on all documents in the collection at $c = 1{,}000$.

and we assign the corresponding documents to the bin. We show the values of documents' retrievability scores that contribute to the wealth of each bin based on the results of the *BM25* model using the Q_s and the Q_a query sets at $c = 100$ (see Figure 10) and $c = 1{,}000$ (see Figure 11), e. g.the first bin based on the Q_s query set at $c = 100$ contains all documents whose retrievability score is between 0 and 7.

2. From each bin, we randomly pick 1,000 documents. Then, we formulate a query from each document, with a randomly chosen length between 3 to 7 terms. Then, terms that formulate the query were picked from the most frequent terms in the document until we get the required length. Stopwords, terms with less than 3 characters or a document frequency less than 2, and terms that occur in more than 25% of the documents in the collection are excluded. Finally, we issue these queries (1,000

Table 38: **Gini Coefficients** for all retrieval models with different values of c; all documents in the collection are used for computing the Gini Coefficient. The retrievability score was computed based on the document version granularity.

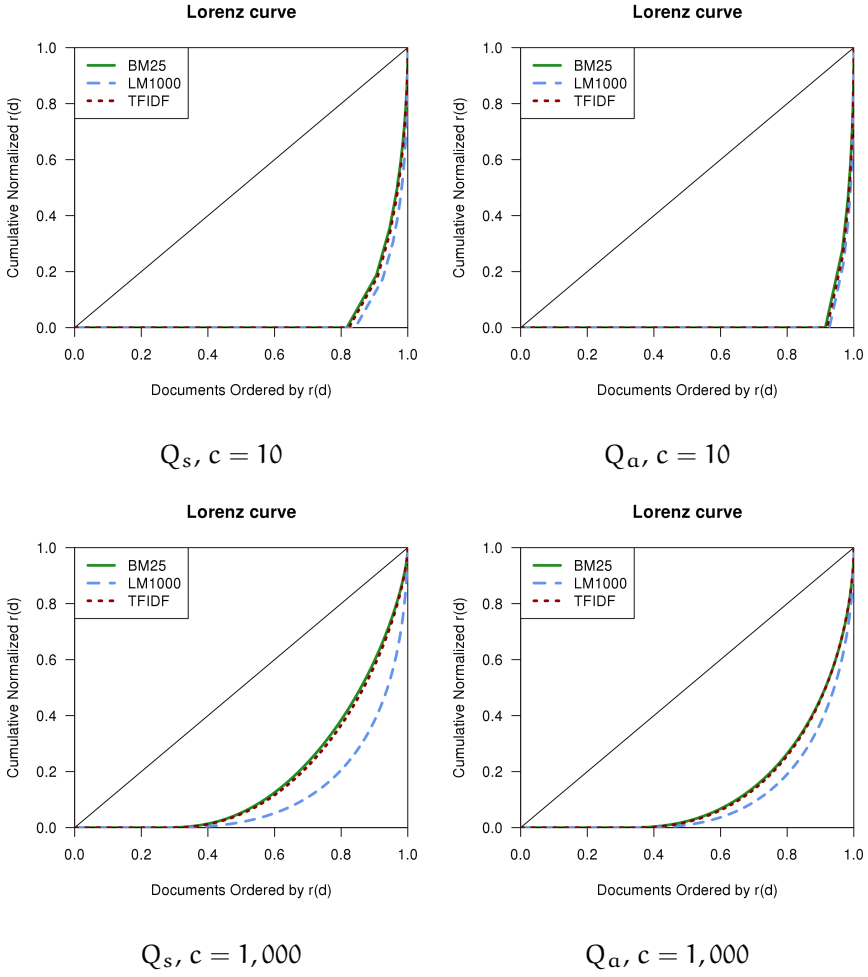
Query Set	Ret. Model	c		
		10	100	1000
Q_a	TFIDF	0.96	0.86	0.73
	BM25	0.95	0.85	0.73
	LM1000	0.96	0.88	0.79
Q_s	TFIDF	0.91	0.78	0.65
	BM25	0.90	0.76	0.63
	LM1000	0.93	0.84	0.77

queries per bin) against the index of the whole collection using *BM25*.

6.5 RETRIEVABILITY BIAS

First, we examine whether the search results obtained using three retrieval models on a Web archive collection are biased **RQ4.1** *Is access to the Web archive collection influenced by a retrievability bias? Can we evaluate and compare retrieval systems on the Web archive collection using the retrievability measure to quantify their retrieval bias?* and investigate the extent of this bias. For this analysis we assumed that a user is looking for an exact version of a document $d_v^{t_i}$. Every document's version was considered as a separate document at indexing time, and thus the relevance granularity was computed at the document's version granularity.

To compare the bias within the different result sets we computed the *Gini Coefficients* using the results of the three models at different cutoff values using the two query sets (Q_a and Q_s described in Section 6.4.1.3) (see Table 38). At $c = 10$, the Gini Coefficients are very high. For example, $G = 0.96$, $G = 0.95$, and $G = 0.96$ for *TF*IDF*, *BM25* and *LM1000*, respectively, based on the Q_a query set. These values are close to total inequality ($G = 1$). For higher values of c, the Gini Coefficients decrease. This trend is the same for the three models using the two query sets. However, even for $c = 1,000$, the Gini Coefficients are still high. The least bias is found in the combination of *BM25* and the Q_s query set at $c = 1,000$ ($G = 0.63$). The largest bias is induced by *LM1000* using the Q_a query set at $c = 10$ ($G = 0.96$). The differences in retrieval bias between the retrieval models, and between different values of c, are visualized in Figure 12. *BM25* induces the smallest inequality for both query sets and can therefore be



$Q_s, c = 10$ $Q_a, c = 10$

$Q_s, c = 1,000$ $Q_a, c = 1,000$

Figure 12: **Retrievability scores inequality** among documents in the **entire collection** visualized with Lorenz Curve.

considered to be the fairest model. This is in line with the findings of [39, 152].

For each setup a number of documents in the collection are never retrieved by any retrieval model ($r(d) = 0$). For the Q_a query set at $c = 10$, only 8% of the documents in the collection were retrieved by *TF*IDF*, 7.3% by *LM1000*, and 8.5% by *BM25*. The large fraction of documents that were *not retrieved* has a strong influence on the high values of the Gini Coefficients. This effect can be seen in the flat line of Lorenz Curves for all c's. For example the Lorenz Curve of *BM25* at $c = 10$ deviates more from the equality line compared to the curve at $c = 1,000$, and has a longer flat line. When only the documents in the union data set were used to assess the bias, the deviation from the equity line was smaller (see Figure 13), all models exhibit less bias compared to the case when all documents in the collection were included computing the bias.

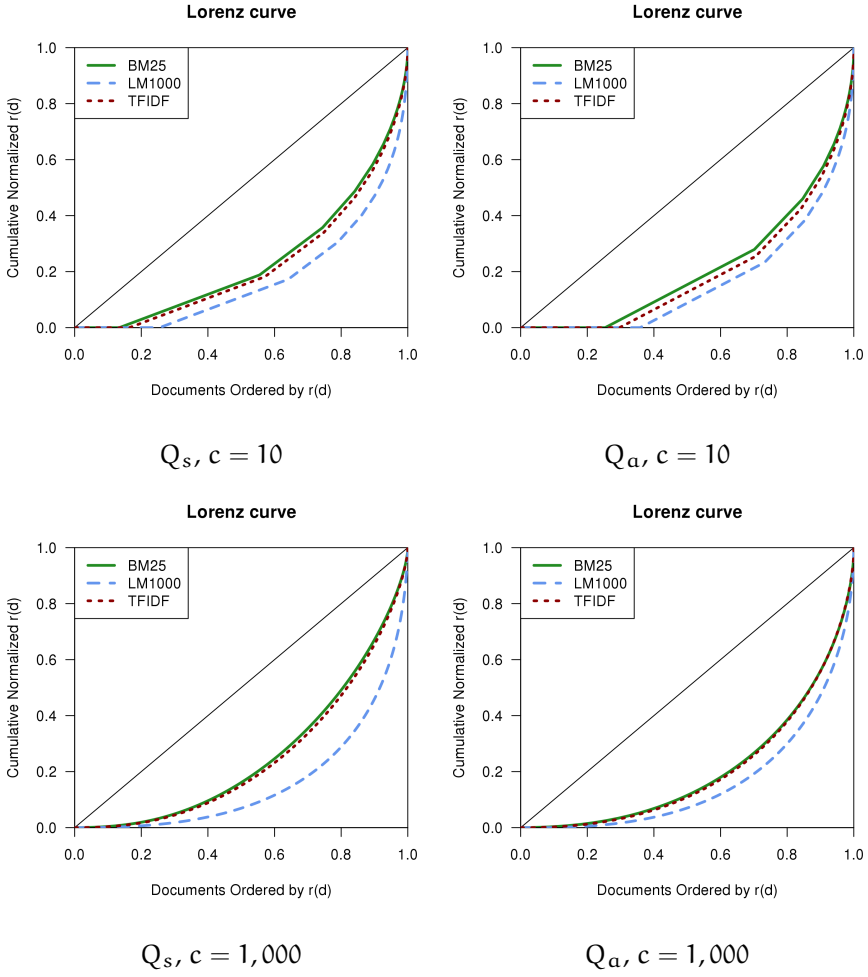


Figure 13: **Retrievability scores inequality** among documents in the ***3Models_union_c*** visualized with Lorenz Curve.

Table 39 shows the Gini Coefficient for all models based on the documents in the *3Models_union_c* set. We cannot directly compare the Gini Coefficient values across the c's as they have been computed with different set sizes. However, we can still compare the models against each other at the same c, for example, we find that the *BM25* model induces the least inequality for both query sets at all values of c.

We are interested in how many of the documents have a chance to be found using the models. The percentage of retrieved documents is the fraction of *unique* documents retrieved using *any* of the three models at a given c to the total number of documents in the collection. As c increases, more documents are retrieved (see Table 39). For example, based on the Q_a query set results, approximately 11% of the documents were retrieved using at least one model at c = 10; the remaining (89%) were *not retrieved* at all. The documents retrieved with *BM25* show the highest overlap with the *3Models_union_c* at dif-

Table 39: **Gini Coefficients** for all retrieval models with different values of c; document versions in the *3Models_union_c* at the corresponding c are considered for computing the Gini Coefficient. The % retrieved is the fraction of retrieved document versions using the models from the whole collection at corresponding c.

Query Set	Ret. Model	c						
		10	20	30	40	50	100	1000
Q_a	TFIDF	0.61	0.62	0.63	0.63	0.63	0.64	0.60
	BM25	0.57	0.59	0.60	0.60	0.60	0.61	0.59
	LM1000	0.67	0.69	0.69	0.70	0.70	0.71	0.68
	% retrieved union	11.4	17.9	22.7	26.4	29.4	39.2	66.3
Q_s	TFIDF	0.55	0.56	0.57	0.57	0.57	0.57	0.51
	BM25	0.53	0.54	0.54	0.55	0.55	0.55	0.49
	LM1000	0.65	0.67	0.68	0.68	0.69	0.70	0.69
	% retrieved union	21.1	30.8	36.6	40.6	43.6	52.0	72.4

ferent c's. For example, considering the *3Models_union_c* set created at $c = 10$, the percentage of overlap between the set of retrieved documents using the *BM25* model and the *3Models_union_c* set equals 75% (for query set Q_a) and 87% (for Q_s). On the other hand, for *LM1000* these percentages equal 64% and 75%, respectively.

6.5.1 *Retrievability and Findability*

We explore the relation between the retrievability score and the findability of a document. We test the hypothesis in [39] which states that the lower the retrievability score of a document, the more difficult it should be to find it, even if the query is tailored to retrieve the target document. We use the known-item search setup as described in Section 6.4.2 to validate this hypothesis.

We computed the Mean Reciprocal Rank (MRR) to measure the effectiveness of the queries from each bin (see Table 40). We compare the MRR distributions of the first three bins with the fourth bin and test whether the differences between the bins are significant using the Kolmogorov-Smirnov test. We found that the bins with higher retrievability scores also have a higher mean MRR score. The largest difference in the MRR distributions is between the first bin and the fourth bin for the two query sets and for both $c = 100$, and $c = 1,000$. Using the Kolmogorov-Smirnov test, we can confirm that it is significantly easier to find documents from the fourth bin compared to documents from the first bin. This confirms our hypothesis and is in line with the findings presented in [42].

Table 40: Effectiveness of known-item queries measured by MRR. The first bin consists of the least retrievable documents, while the fourth bin contains the most retrievable documents. An $*$ indicates that the difference between the corresponding bin and the fourth bin is not significant using the Kolmogorov-Smirnov ($p > 0.05$).

Query Set	Rank cutoff c	Bins			
		1st	2nd	3rd	4th
Q_a	$c = 100$	0.12	0.35	0.37*	0.40
	$c = 1000$	0.12	0.25	0.25	0.31
Q_s	$c = 100$	0.09	0.30*	0.31*	0.30
	$c = 1000$	0.07	0.23*	0.25*	0.24

6.6 IMPACT OF NUMBER OF VERSIONS ON THE RETRIEVABILITY BIAS

In Section 6.5, we showed that all retrieval models impose a retrievability bias on the Web archive collection when we use the document's version as the basis. In this section, we explore the effect of varying numbers of versions of the same document on the retrievability bias **RQ4.2** *How does the number of versions of documents in the Web archive collection influence the retrievability bias of a retrieval system?*. First, we show how collapsing similar versions of the same document based on content similarity influences the retrieval bias (Section 6.6.1). Then, we use the number of versions per document to refine the search results after linearly combining a prior based on the number of versions with a score given using the retrieval model. In this approach, we collapse versions of the same documents based on their URLs (Section 6.6.2).

6.6.1 *Collapsing Similar Versions*

We first consider as a successful retrieval when the system returns any version of a specific document. In this scenario, the retrievability score of a document is computed by aggregating the retrievability scores of its versions. In a second scenario, we take the view that the content of the document's versions may have changed over time. Therefore, we cluster versions of the same document based on the similarity of their content, and we aggregate the retrievability scores at the cluster level. We believe that this experiment can be helpful when deciding which version(s) of a document to show to the user in the result lists as it allows other documents to appear in the top of the ranked results. We base the following experiments on the document's versions retrieved using the models, and using the two query sets (discussed in Section 6.5).

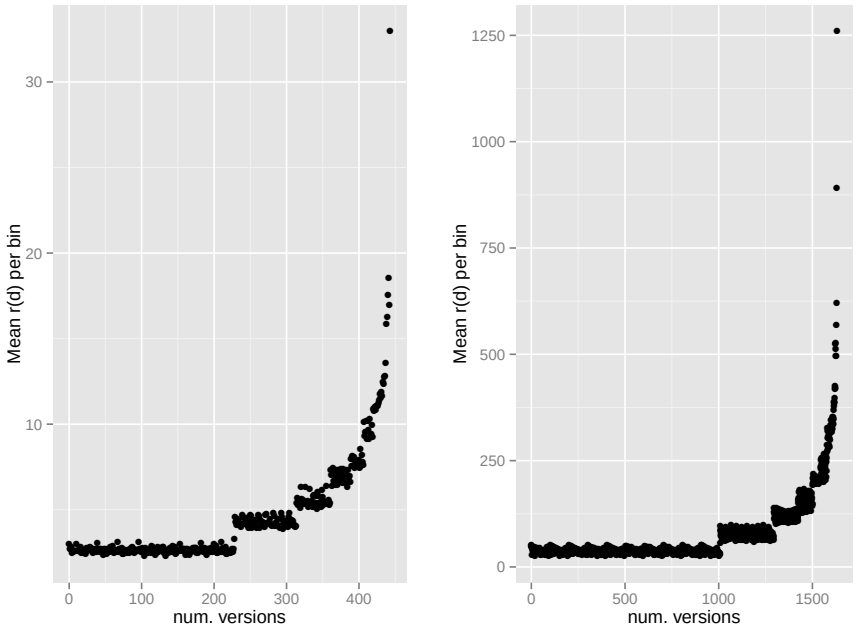
Table 41: **Gini Coefficients** for all retrieval models based on the two query sets. **Any version**.

Query Set	Ret. Model	c						
		10	20	30	40	50	100	1000
Q_a	TFIDF	0.68	0.70	0.70	0.71	0.71	0.71	0.69
	BM25	0.66	0.67	0.68	0.68	0.69	0.69	0.69
	LM1000	0.74	0.75	0.75	0.76	0.76	0.76	0.75
	% retrieved union	11.7	17.3	21.4	24.7	27.3	36.0	62.5
Q_s	TFIDF	0.64	0.65	0.66	0.66	0.66	0.67	0.61
	BM25	0.62	0.63	0.64	0.65	0.65	0.65	0.60
	LM1000	0.71	0.73	0.74	0.74	0.75	0.75	0.73
	% retrieved union	21.2	29.3	34.2	37.7	40.3	48.0	68.9

6.6.1.1 *Any version*

In this experiment, we consider finding any version of a document d at a given c a success. We compute the retrievability score $r(d)$ of a document d by accumulating the retrievability scores of its versions $r(d_v^{t_i})$. In the previous section, the retrievability scores were computed for document versions. In order to compute the retrievability score for documents, we map every document's version identifier to its URL. After that, we compute the Gini Coefficients for the three models with different c based on the documents in the union (see Table 41).

We found that the aggregation at document level increases the inequality bias for all retrieval models at all c's for the two query sets. We can derive that from the comparison of the Gini Coefficients in Table 41 with those in Table 39 (when the retrievability scores were computed at document version granularity). This can be explained by the varying number of versions per URL. On average every document is represented with 1.8 versions in the collection (see Table 34). Documents with a higher number of versions obtain higher retrievability scores as their versions are likely to appear multiple times in the ranked results at a given c. A similar trend exists for the other models, and also for the Q_s query set. In order to find out whether documents with a higher number of versions obtain higher retrievability scores, we plot the number of versions vs. retrievability scores. We did that by first sorting documents based on their number of versions and dividing them into bins. Each bin consists of $20,000$ documents. For each bin, we calculated the mean retrievability score. We found that as the number of versions increases, the retrievability score increases as well (see Figure 14).



Q_s, $c = 10$ Q_s, $c = 1000$

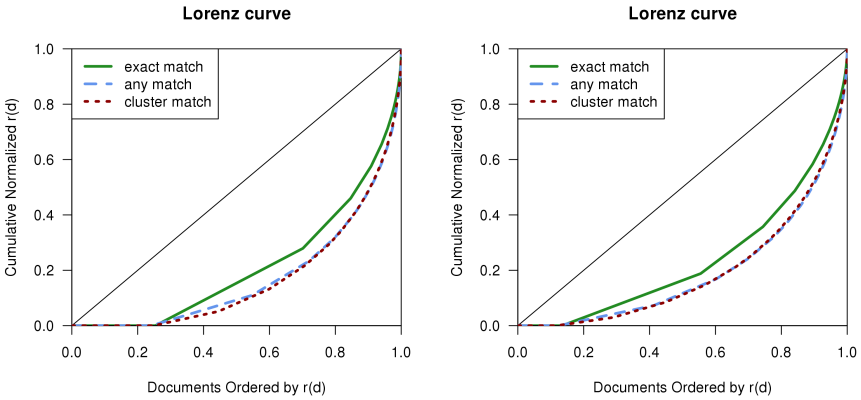
Figure 14: Number of versions vs. retrievability score, for the *BM25* model.

6.6.1.2 *Clustering Versions (Content-based Similarity)*

In the previous experiment, we showed that the inequality increases when the $r(d)$ was computed at document (URL) granularity, by aggregating the retrievability score of all versions of the same document. As a next step, we explore the effect of grouping the most similar versions of the same document into two clusters. For every document in the Web archive collection, we first collect all versions of that document. We create a term frequency vector for each version and compute the cosine similarity between the versions. Finally, we split them into two clusters based on their similarity. We modify the retrieved results by the models by replacing the document's version identifier with the corresponding cluster identifier. Based on the mapping between document's version and cluster IDs, we compute the retrievability scores for every cluster.

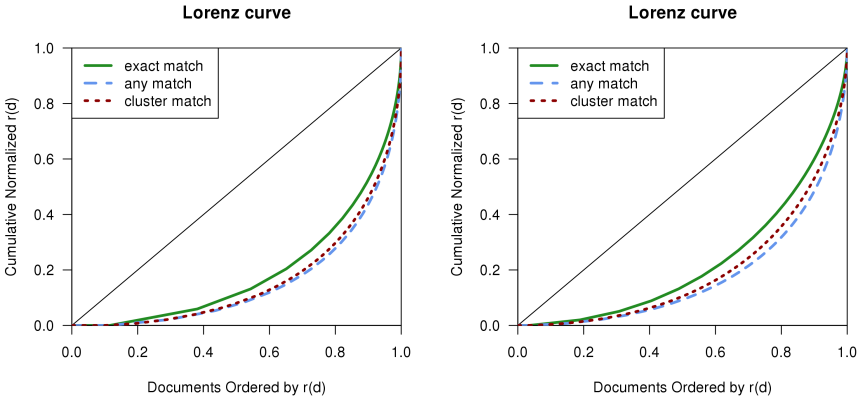
Table 42 shows the Gini Coefficient for all retrieval models based on the Q_a and the Q_s query sets. Comparing the Gini Coefficients with those in Table 41 shows that the bias is smaller in the case of clustering compared to the *any version* case. Also the percentage of retrieved cluster IDs in the union of all models at given c is higher than the percentage of retrieved versions in the union at the corresponding c.

The Lorenz Curves show that the least bias is found when the retrievability score was computed at the document's version level (see Figure 15). The bias increases when the retrievability score was computed at the document's level considering the two scenarios; the red



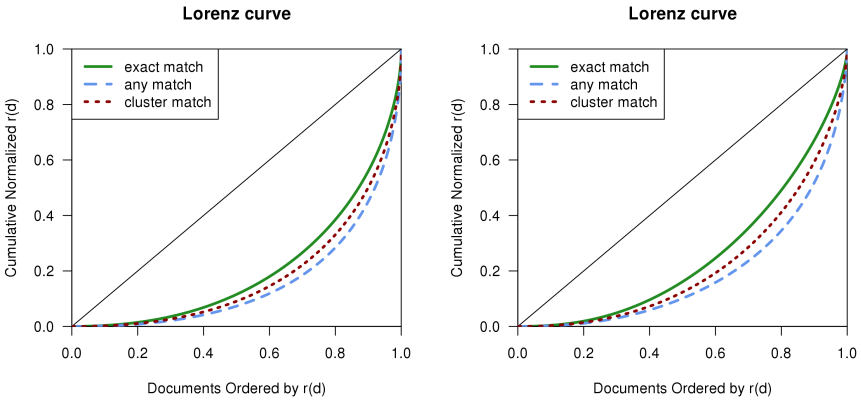
Q_a, $c = 10$

Q_s, $c = 10$



Q_a, $c = 100$

Q_s, $c = 100$



Q_a, $c = 1{,}000$

Q_s, $c = 1{,}000$

Figure 15: Lorenz curves visualizing the inequality of retrievability scores induced by *BM25* for three scenarios; exact match (green), any match (blue), and cluster match (red).

Table 42: **Gini Coefficients** for all retrieval models based on the two query sets. **Cluster version**.

Query Set	Ret. Model	c						
		10	20	30	40	50	100	1000
Q_a	TFIDF	0.69	0.70	0.70	0.70	0.70	0.70	0.65
	BM25	0.67	0.67	0.68	0.68	0.68	0.68	0.65
	LM1000	0.75	0.76	0.76	0.76	0.76	0.76	0.72
	% retrieved union	18.6	27.5	33.7	38.5	42.3	54.0	81.5
Q_s	TFIDF	0.64	0.64	0.64	0.64	0.64	0.64	0.58
	BM25	0.62	0.62	0.62	0.62	0.62	0.62	0.57
	LM1000	0.72	0.73	0.74	0.74	0.74	0.74	0.72
	% retrieved union	34.2	45.8	52.4	56.7	59.9	68.5	87.3

and the blue curves are more deviated from the equality line. The bias is less based on the clustering of similar versions (red curve) compared to any match (blue curve); the difference is bigger at higher c.

6.6.2 *Collapsing Versions (URL-based)*

We showed that multiple versions of the same document impact the retrievability bias. This bias was the highest when the retrieval granularity was the document's version. In this section, we investigate the change in the retrieval bias when all versions of the same document are merged into one entry in the search result list based on their URLs. However, we take the number of versions into account for ranking documents, by embedding a prior based on the number of versions, with the retrieval models.

When a query q is issued, the retrieval model is used in computing a score (IR_{score}) for each document d in the collection based on how relevant its content is to the query q. Then the documents are ranked based on their relevance scores.

Including the temporal aspect of Web archives into retrieval models was discussed in [69]. In their model, they linearly combined a prior which favors documents with more versions or longer existence (time span between first version and last version) with known IR models. They showed that this approach achieved significant improvement over the baseline IR model.

We copy their approach and linearly combine the relevance score given to a document using a retrieval model (IR_{score}) with a score

based on the number of versions for that document using the following formula:

$$IR_{score}^{versions} = \lambda * IR_{score} + (1 - \lambda) * prior_{versions} \quad (6)$$

where IR_{score} is the relevance score as computed using the retrieval model for a document d and a given query q. The $prior_{versions}$ is a prior based on the number of versions; this prior is independent from the retrieval model. The value of this prior increases with the number of versions and is computed as follows:

$$prior_{versions} = \frac{\log_{10}(\#Versions)}{\log_{10}(max.\#Versions)} \quad (7)$$

The number of versions per document is divided by the maximum $\log(max.\#Versions)$ in order to normalize the values to range from 0 to 1. We also normalize the values of IR_{score} given using the models to the same range. The retrieved documents are ranked from 1 to 1,000 using the retrieval model for each query, and every document is assigned a score (IR_{score}). If the same document appears multiple times, then we take the maximum score. We adjusted the search results for each query by computing and sorting documents based on the new scores using Equation 6. Finally, we computed the retrievability score using the documents in *3Models_union_c*.

We compared the Gini Coefficients of this experiment (Table 43) with results obtained by accumulating the retrievability scores of versions of the same document (Section 6.6, Table 41). We found that the inequality decreases for all models at all c's. This means that collapsing the versions of the same document reduces the retrievability bias induced by all models. However, the bias is still high, with the Gini Coefficient in the range between 0.51 and 0.75.

The percentage of retrieved documents increases because the retrieved items in the search results are the documents instead of the document's versions. We see a similar pattern for all values of c until we reach 1,000; the percentage decreases as it approaches the maximum number of documents retrieved for the query. The difference in percentage retrieved in this experiment and the *any match* case increases as c increases.

6.7 QUANTIFICATION OF RETRIEVAL BIAS OVER THE YEARS

We investigated how the bias imposed by the retrieval system correlates with the number of documents aggregated over the years **RQ4.3** *Does a retrieval system favor specific subsets of the collection?*. The Web archive collection consists of several crawls accumulated over time, and the number of websites included in the crawling process increased over the years. Therefore, the number of crawled documents

Table 43: **Gini Coefficients** for the three retrieval models based on the two query sets, after embedding the prior based on number of versions with content similarity weight.

Query Set	Ret. Model	c						
		10	20	30	40	50	100	1000
Q_a	TFIDF	0.64	0.65	0.66	0.66	0.67	0.67	0.62
	BM25	0.62	0.63	0.64	0.65	0.65	0.66	0.61
	LM1000	0.73	0.74	0.75	0.75	0.75	0.75	0.70
	% retrieved union	13.8	20.4	24.8	28.2	30.9	39.7	60.9
Q_s	TFIDF	0.60	0.61	0.62	0.62	0.62	0.62	0.53
	BM25	0.58	0.60	0.60	0.61	0.61	0.61	0.51
	LM1000	0.70	0.72	0.73	0.73	0.74	0.74	0.70
	% retrieved union	24.5	33.0	38.1	41.6	44.2	51.8	66.6

varies. We explore whether the number of documents crawled in one year has an impact on the number of documents retrieved. For this experiment, we focused on *BM25* as it induced the smallest bias (see Section 6.5).

As mentioned in Section 6.4.1.1, every document's version in the Web archive collection has an associated crawling timestamp. We used this timestamp to divide the retrieved documents according to the year in which they were archived. This led to four subsets, 2009, 2010, 2011, and 2012. We apply the time-based splitting using the retrievability scores of documents computed for *BM25*, using the two query sets at different values of c: $c = 10$, 100, and $1,000$.

For every subset, we computed the mean retrievability score. We did not find a relation between mean retrievability score and subset size (see Table 44). The result is in line with [39]; for subsetting based on website domains, they found that there is no relation between subset size and the mean retrievability score computed per domain subset. As expected, we found a relation between subset size and the percentage of retrieved documents. The larger a subset is, the higher the percentage of retrieved documents. For every subset, we computed the fraction of retrieved documents at given c, where the subset size is the same for all c's. The percentage of retrieved documents increases over the years until 2011, then drops for 2012 (see Table 44). We can explain this behavior by the number of documents that were crawled in each year. For example, the largest number of documents was crawled in 2011, and the highest percentage retrieved using *BM25* at all c's is from that same year.

Table 44: **Retrievability subset analysis** using *BM25* results. For every subset, query set, and c: We present the fraction of retrieved documents from the subset in percentage (num. retrieved / subset size) (**first column**). The mean retrievability score of retrieved documents (**second column**). The fraction of retrieved documents from the corresponding subset to the number of all retrieved (num. retrieved per subset / all retrieved (all subsets)) in percentages (**third column**); sum of the percentage in this column is equal to 100%.

	Q: Q_a								
subset (size)	$c = 10$			$c = 100$			$c = 1,000$		
2009 (12,232,831)	5.5	2.2	9.4	23.6	4.8	9.8	47.3	23.4	10.5
2010 (22,596,291)	7.9	2.1	24.7	33.0	5.0	25.4	63.0	25.2	25.9
2011 (30,275,150)	8.9	2.1	37.3	37.0	5.0	38.1	68.4	25.9	37.7
2012 (19,464,431)	10.6	2.2	28.6	40.4	5.4	26.7	73.1	28.1	25.9

	Q: Q_s								
subset (size)	$c = 10$			$c = 100$			$c = 1,000$		
2009 (12,232,831)	12.5	2.7	9.9	34.9	7.9	10.0	53.8	39.4	10.8
2010 (22,596,291)	17.3	2.6	25.2	48.6	7.8	25.8	71.4	39.5	26.5
2011 (30,275,150)	19.3	2.5	37.7	53.6	7.9	38.2	75.1	41.1	37.4
2012 (19,464,431)	21.7	2.6	27.3	56.9	8.3	26.0	79.0	43.8	25.3

6.7.1 *Time-based Subsets based on Time-based Queries*

By binning the retrieved documents by year, we showed that the percentage of retrieved documents from a particular subset correlates with the number of documents in the bin. This analysis was based on simulated queries. Therefore, the number of queries we extracted from one year is directly linked to the number of documents that were crawled in that same year.

We further explore the relation between the queries' timestamps and the documents' timestamps. We focused our analysis on Q_a because the Anchor Text is known to be a good substitute for both documents' titles and real queries. Recall that in Section 6.4.1.3, we generated the Q_a query set with a timestamp for each query which represents the crawling date. We divided the queries into 4 subsets, one for each year. We refer to these query sets as Q_a *_YYYY*, e.g., Q_a *_2009* represents Anchor Text extracted from links that were extracted from pages crawled in 2009. The number of Anchor Text increases over years (see Table 46), but then drops for 2009. Because some documents exist in multiple versions, we expected to have overlapping Anchor Text across the subsets. Therefore, along with the number of queries, we also show the number of unique queries per year compared to all previous years (see Table 46). For example, 41.9% of Anchor Text from 2012 are new; they did not exist in any year before.

Table 45: Query length distribution in the Q_a query set per year.

query length	2009	2010	2011	2012
1	26.2	23.2	23.7	23.9
2	33.5	34.9	34.1	34.4
3	24.2	24.9	24.9	23.5
4	11.8	12.7	12.9	13.0
5	3.8	3.8	3.9	4.7
6	0.5	0.5	0.5	0.5

The average query length is almost the same for all subsets (see Table 46). The distribution of query lengths is the same over the years (see Table 45).

Q_a *_2011* has the highest vocabulary size (see Table 46). The number of queries in 2012 is less than the number of queries in 2011 because fewer documents were crawled in 2012 compared to 2011 (see Table 34). In total, there are 100,908 terms shared across the vocabulary of the four query subsets.

We repeated the retrievability assessment as discussed in Section 6.4.1.4, with the four query subsets. We issued every $q \in Q_a$ *_YYYY* for all subsets against the index of the entire collection. The query subsets are generated from the Q_a query set. Therefore, in order to explore the influence of these query subsets on the bias, we used the documents in the *3Models_union_c* set and the Q_a query set. When we compare the Gini Coefficients for the three retrieval models, we see that *BM25* leads to the smallest bias for the four query subsets at all of the studied values of c (see Table 47). The percentage of retrieved documents has an effect on the extent of the retrieval bias for all retrieval models. For example, the Q_a *_2009* query set shows the highest inequality for all retrieval systems because it has the smallest percentage of retrieved documents, whereas the Q_a *_2011* shows the smallest bias and has the highest percentage of retrieved documents. The result of this experiment confirms a relation between retrieval bias and number of documents crawled per year. We further investigate the relation between the timestamps of the queries and the timestamps of retrieved documents.

We performed a subset analysis based on the documents retrieved with *BM25* using the four query subsets, to measure differences in the retrieval bias over the years. For example, using the timestamps of documents retrieved with the *BM25* model using the Q_a *_2009* query subset, we partitioned the documents into 4 subsets, at different c's. For each subset and c we computed the mean retrievability score and the percentage of documents in that subset relative to the total, as we did in the subset analysis based on the Q_a. In addition to that, we computed the relative increase in the fraction of retrieved

Table 46: Summary of query subsets of Q_a query set. For each subset, we show the number of queries. In parentheses is the number of unique queries in the corresponding subset (year) compared to previous years. For example, the Q_a *_2012* is compared against 2009, 2010, and 2011. For the 2009 subset the percentage of unique Anchor Text is *N/A* as it is the first, and the percentage decreases across the years.

Q_a subsets	# of queries	Mean (query length)	#of terms
Q_a *_2009*	358,745 (N/A)	2.3	201,198
Q_a *_2010*	664,678 (69.0%)	2.4	326,725
Q_a *_2011*	998,350 (59.1%)	2.4	475,590
Q_a *_2012*	848,999 (41.9%)	2.4	411,263

documents compared to running the Q_a query set (Table 44). This gives us an indication of how many documents we can retrieve from 2009 by running 2009 queries (Q_a *_2009*) compared to those we get by running queries from all years (Q_a). Running queries from a particular year causes the highest increase in the fraction of retrieved documents from that year (see Table 48). There is a relation between the timestamp of the queries and the timestamps of the documents. For example, using Q_a *_2009* at $c = 10$, 14.2% with retrieved documents using *BM25* originated from 2009, while by using all Anchor Text from all years (Q_a) at the same c, 9.4% of retrieved documents were from 2009. Running 2009 queries therefore results in a +4.8% increase of documents retrieved from that year. However, this effect decreases for higher *c*'s.

6.8 DISCUSSION & CONCLUSIONS

In Web archives, the main focus has been on preserving the content from the Web before it is lost. Recently, Web archive initiatives started to make their Web archive collections available for search through full-text search systems, so as of yet, there are not many studies into the evaluation of Web archive search systems. The lack of queries with judged relevant documents for web archives complicates such research. Retrievability has been proposed as an alternative that does not require relevance assessment, a measure that allows to quantify accessibility bias. Retrievability has been applied in various studies on community-collected test collections such as the TREC collections. The documents in Web archives differ however from those in previously studied collections, because they are typically available in multiple versions which can be an implicit source of bias. We used the retrievability score per document and the overall bias measured by the Gini Coefficient and the Lorenz Curve of the retrievability scores

Table 47: **Gini Coefficients** for the three models at different c's using different query subsets, using documents in the *3Models_union_c* generated based on running the Q_a query set.

Query Set	Ret. Model	c						
		10	20	30	40	50	100	1000
Q_a *_2009*	TFIDF	0.85	0.84	0.82	0.81	0.81	0.77	0.64
	BM25	0.84	0.83	0.81	0.80	0.79	0.76	0.64
	LM1000	0.88	0.87	0.86	0.85	0.84	0.82	0.72
	% retrieved union	3.7	6.6	9.0	11.2	13.2	20.8	56.4
Q_a *_2010*	TFIDF	0.76	0.75	0.74	0.73	0.72	0.70	0.60
	BM25	0.74	0.73	0.72	0.71	0.70	0.68	0.60
	LM1000	0.80	0.79	0.78	0.78	0.77	0.76	0.68
	% retrieved union	6.2	10.5	14.0	17.0	19.5	28.8	62.0
Q_a *_2011*	TFIDF	0.70	0.69	0.69	0.68	0.68	0.67	0.60
	BM25	0.67	0.67	0.66	0.66	0.66	0.65	0.59
	LM1000	0.74	0.74	0.74	0.74	0.74	0.73	0.68
	% retrieved union	8.1	13.4	17.5	20.8	23.6	33.3	64.0
Q_a *_2012*	TFIDF	0.72	0.71	0.70	0.69	0.69	0.67	0.59
	BM25	0.69	0.68	0.68	0.67	0.67	0.65	0.59
	LM1000	0.76	0.76	0.75	0.75	0.75	0.74	0.68
	% retrieved union	7.4	12.4	16.2	19.5	22.2	31.8	63.4

of all documents to quantify the overall bias imposed by the retrieval model on the collection. We measured the retrievability and the overall bias in different scenarios in order to check how the retrievability measure behaves under different retrieval models and different search scenarios. We also investigated whether search results in Web archives are influenced by varying number of versions, and how retrieval systems that are adapted to deal with them can be evaluated using retrievability.

We assessed the retrievability bias induced by three retrieval systems using retrievability scores, which we computed for each document's version in the collection. Our results show that the three systems induce bias at a document's version level, and there is a relation between the retrievability score of a document and the difficulty level of finding that document. Documents with higher retrievability scores are significantly easier to find, thus confirming that the retrievability score is a useful metric.

Table 48: **Retrievability subset analysis based on time-aware queries** using *BM25* results. The fraction of retrieved documents per year to the total documents retrieved using *BM25* (*%retrieved*). The *%gain* represents the relative percentage of documents that we get per year using the corresponding query set to the % retrieved of the same year using the entire Q_a query set (Table 44).

	Q_a *_2009*		Q_a *_2010*		Q_a *_2011*		Q_a *_2012*	
	Mean r(d)	%retrieved (%gain)	Mean r(d)	%retrieved (%gain)	Mean r(d)	%retrieved (%gain)	Mean r(d)	%retrieved (%gain)
	$c = 10$							
2009	1.8	14.2 **(+4.8)**	1.7	9.9 (+0.5)	1.7	8.9 (-0.5)	1.6	8.7 (-0.7)
2010	1.5	26.0 (+1.3)	1.8	28.3 **(+3.5)**	1.7	23.9 (-0.8)	1.6	22.7 (-2.0)
2011	1.4	34.3 (-3.0)	1.6	35.8 (-1.5)	1.9	39.6 **(+2.3)**	1.7	36.4 (-0.9)
2012	1.4	25.5 (-3.0)	1.6	26.1 (-2.5)	1.8	27.6 (-1.0)	1.9	32.1 **(+3.5)**
	$c = 100$							
2009	2.7	11.4 **(+1.6)**	2.9	9.8 (0.0)	3.3	9.5 (-0.3)	2.9	9.4 (-0.4)
2010	2.3	25.9 (+0.5)	3.1	26.4 **(+1.0)**	3.4	25.1 (-0.3)	3	24.7 (-0.7)
2011	2.1	36.8 (-1.3)	2.7	37.5 (-0.5)	3.7	38.6 **(+0.6)**	3.1	37.9 (-0.2)
2012	2.2	26.0 (-0.8)	2.9	26.2 (-0.5)	3.6	26.7 (0.0)	3.6	27.9 **(+1.2)**
	$c = 1,000$							
2009	7.4	10.8 **(+0.3)**	10.5	10.5 (0.0)	13.6	10.5 (-0.1)	11.7	10.4 (-0.1)
2010	6.7	25.7 (-0.2)	11.5	25.9 (0.0)	14.8	25.8 (-0.1)	12.6	25.7 (-0.2)
2011	6.4	37.4 (-0.3)	10.7	37.6 (-0.1)	16.1	37.8 **(+0.1)**	13.3	37.7 (0.0)
2012	6.7	26.1 (+0.2)	11.3	26.0 **(+0.1)**	16.2	26.0 **(+0.1)**	15.5	26.2 **(+0.3)**

Then, we studied the change in bias when the system is adapted to deal with multiple versions of a document. We explored this using two approaches to collapse versions of the same document. First, we collapse document's versions based on their content similarity (*clustering-based*). Here, the cluster with more versions will get a higher retrievability score. Second, we collapse the versions based on their URL. Here, we embed a prior (based on the number of versions) with the scores given by retrieval systems; this means a document with more versions gets a higher score. The *clustering-based* approach takes into account that the content of document's versions may change over time, and thus collapses them into clusters. The *URL-based* approach considers them similar and collapse them into one URL. The bias was lower for the two collapsing approaches, as compared with the systems which do not consider the multiple versions of the document. The three retrieval systems impose lower bias in the URL approach, as compared to the clustering approach. We have shown that retrievability is suitable to assess Web archive retrieval systems, by showing its ability to capture the bias based on the approach followed to deal with multiple versions.

The evaluation of Web archives in terms of accessibility is important for both the institutions maintaining the archives and the users searching the archive. Knowing which documents are particularly hard to find allows the institutions to improve their retrieval systems and the users to adapt their search strategies and be aware of the retrieval bias and the source of that bias.

Part II

OPEN WEB (LIVE AND DYNAMIC) & CRAWLED WEB (ARCHIVED AND STATIC)

7

THE STRANGE CASE OF REPRODUCIBILITY VS. REPRESENTATIVENESS IN CONTEXTUAL SUGGESTION TEST COLLECTIONS

The most common approach to measuring the effectiveness of Information Retrieval systems is by using test collections. The Contextual Suggestion (CS) TREC track provides an evaluation framework for systems that recommend items to users given their geographical context. The specific nature of this track allows the participating teams to identify candidate documents either from the *Open Web* or from the *ClueWeb12* collection, a static version of the web. In the judging pool, the documents from the *Open Web* and *ClueWeb12* collection are distinguished. Hence, each system submission should be based only on one resource, either *Open Web* (identified by URLs) or *ClueWeb12* (identified by ids). To achieve reproducibility, ranking web pages from *ClueWeb12* should be the preferred method for scientific evaluation of contextual suggestion systems, but it has been found that the systems that build their suggestion algorithms on top of input taken from the *Open Web* achieve consistently a higher effectiveness. Because most of the systems take a rather similar approach to making contextual suggestions, this raises the question whether systems built by researchers on top of *ClueWeb12* are still representative of those that would work directly on industry-strength web search engines. Do we need to sacrifice reproducibility for the sake of representativeness?

We study the difference in effectiveness between *Open Web* systems and *ClueWeb12* systems through analyzing the relevance assessments of documents identified from both the *Open Web* and *ClueWeb12*. Then, we identify documents that overlap between the relevance assessments of the *Open Web* and *ClueWeb12*, observing a dependency between relevance assessments and the source of the document being taken from the *Open Web* or from *ClueWeb12*. After that, we identify documents from the relevance assessments of the *Open Web* which exist in the *ClueWeb12* collection but do not exist in the *ClueWeb12* relevance assessments. We use these documents to expand the *ClueWeb12* relevance assessments.

Our main findings are twofold. First, our empirical analysis of the relevance assessments of two years of CS track shows that *Open Web* documents receive better ratings than *ClueWeb12* documents, especially if we look at the documents in the overlap. Second, based on an expanded version of the relevance assessments and on generating *ClueWeb12*-based runs from *Open Web* runs, we have investi-

gated the representativeness of *ClueWeb12* collection. Although the performance of *Open Web* systems decreases, we find a representative sample of *ClueWeb12* collection in *Open Web* runs.

7.1 INTRODUCTION

Recommender systems aim to help people find items of interest from a large pool of potentially interesting items. The users' preferences may change depending on their current context, such as the time of the day, the device they use, or their location. Hence, those recommendations or suggestions should be tailored to the context of the user. Typically, recommender systems suggest a list of items based on users' preferences. However, awareness of the importance of context as a third dimension beyond users and items has increased, for recommendation [32] and search [131] alike. The goal is to anticipate users' context without asking them, as stated in The Second Strategic Workshop on Information Retrieval (`SWIRL 2012`) [34]: "Future information retrieval systems must anticipate user needs and respond with information appropriate to the current context without the user having to enter a query". This problem is known as `contextual suggestion` in Information Retrieval (`IR`) and *context-aware recommendation* in the Recommender Systems (`RS`) community.

The TREC Contextual Suggestion (CS) track introduced in 2012 provides a common evaluation framework for investigating this task [79]. The aim of the CS task is to provide a list of ranked suggestions, given a location as the (current) user context and past preferences as the user profile. The public *Open Web* was the only source for collecting candidate documents in 2012. Using APIs based on the *Open Web* (either for search or recommendation) has the disadvantage that the end-to-end contextual suggestion process cannot be examined in all detail, and that reproducibility of results is at risk [99, 98]. To address this problem, starting from 2013 participating teams were allowed to collect candidate documents either from *Open Web* or from the *ClueWeb12* collection.

In the 2013 and 2014 editions of CS track, there were more submissions based on the *Open Web* compared to those based on the *ClueWeb12* collection. However, to achieve reproducibility, ranking web pages from *ClueWeb12* should be the preferred method for scientific evaluation of contextual suggestion systems. It has been found that the systems that build their suggestion algorithms on top of input taken from the *Open Web* achieve consistently a higher effectiveness than systems based on the *ClueWeb12* collection. Most of the existing works have relied on public tourist APIs to address the contextual suggestion problem. These tourist sites (such as Yelp and Foursquare) are specialized in providing tourist suggestions, hence those works are focused on re-ranking the resulting candidate suggestions based

on user preferences. Gathering suggestions (potential venues) from the *ClueWeb12* collection has indeed proven a challenging task. First, suggestions have to be selected from a very large collection. Second, these documents should be geographically relevant (the attraction should be located as close as possible to the target context), and they should be of interest for the user.

The finding that Open Web results achieve higher effectiveness raises the question whether research systems built on top of the *ClueWeb12* collection are still representative of those that would work directly on industry-strength location-based search engines. We focus on analyzing reproducibility and representativeness of the *Open Web* and *ClueWeb12* systems. We study the gap in effectiveness between *Open Web* and *ClueWeb12* systems through analyzing the relevance assessments of documents returned by them. After that, we identify documents that overlap between *Open Web* and *ClueWeb12* results. We define two different sets of overlap: First, the overlap in the relevance assessments of documents returned by *Open Web* and *ClueWeb12* systems, to investigate how these documents were judged according to the relevance assessments gathered when they were considered by *Open Web* or *ClueWeb12* systems. The second type of overlap is defined by the documents in the relevance assessments of the *Open Web* systems which are in *ClueWeb12* collection but not in the relevance assessments of *ClueWeb12* systems. The purpose is to use the judgments of these documents (mapped from *Open Web* on *ClueWeb12* collection) to expand the relevance assessments of *ClueWeb12* systems resulting on having a new test collection. Figure 16 illustrates these different test collections, the details given in Section 7.3.3. Then, we focus on how many of the documents returned by *Open Web* systems can be found in the *ClueWeb12* collection, an analysis to assess the reproducibility point of view. Finally, we apply the knowledge about the tourist information available in the *Open Web* for selecting documents from *ClueWeb12* to find a representative sample from the *ClueWeb12* collection. Specifically, we address the following research questions:

RQ5 *Do relevance assessments of Open Web differ (significantly) from relevance assessments of ClueWeb12 documents? Can we identify an overlap between the two sets, and the documents in the overlap were judged?*
We divide this research question into the following sub-research questions:

RQ5.1 *Do relevance assessments of Open Web differ (significantly) from relevance assessments of ClueWeb12 documents?*

RQ5.2 *Can we identify an overlap between Open Web systems and ClueWeb12 systems in terms of documents suggested by both?, how are those documents in the overlap judged?*

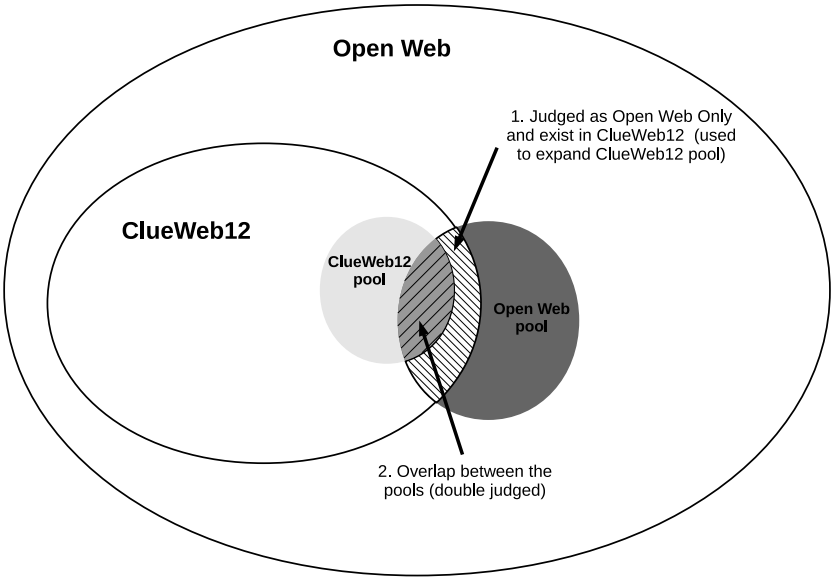


Figure 16: Illustration of the relation between pools and the source of the documents. Subset 1 represents the documents in the *Open Web* pool and were found in *ClueWeb12* collection but do not exist in the *ClueWeb12* pool (this subset is used to expand the *ClueWeb12* pool). Subset 2 represents the overlap between the *Open Web* pool and *ClueWeb12* pool, documents in this subset were double judged (we use this subset to show the bias between *Open Web* and *ClueWeb12* results).

RQ5.3 *How many of the documents returned by Open Web systems can be found in the ClueWeb12 collection as a whole?*

The remainder of the chapter is organized as follows: first we discuss related work (Section 7.2), followed by a description of the experimental setup (Section 7.3). After that we present an analysis to compare *Open Web* and *ClueWeb12* relevance assessments (Section 7.4). Then we discuss how much of the *Open Web* systems can be reproduced from the *ClueWeb12* collection, and we evaluate them on the *ClueWeb12* test collection (Section 7.5). Finally, we discuss conclusions drawn from our findings (Sections 7.6).

7.2 RELATED WORK

In the Recommender Systems area, recommendation algorithms for several types of content have been studied (movies, tourist attractions, news, friends, etc.). These types of algorithms are typically categorized according to the information they exploit: collaborative filtering (based on the preferences of like-minded users [143]) and content-based filtering (based on similar items to those liked by the user [126]). In the Information Retrieval area, approaches to *contextual suggestion* usually follow a content-based recommendation approach.

The majority of related work results from the corresponding TREC track, focusing on the specific problem of how to provide tourist attractions given a location as context, where many participants have relied on APIs of location-based services on the *Open Web*. Candidate suggestions based on location are then ranked based on their similarity with the known user interests. In this case, the key challenge is to model user interests.

Given the description of a set of examples (suggestions) judged by the user, existing studies exploit the descriptions of the suggestions to build her profile, usually represented as the textual information contained in the description of the suggestions. [147] build two user profiles: a positive profile represents terms from those suggestions liked by the user before, whereas a negative profile is based on descriptions of suggestions disliked by the user. In [105, 159] both the descriptions and the categories of the suggestions are used to build the user profiles. In [160], the authors proposed an opinion-based approach to model user profiles by leveraging similar user opinions of suggestions on public tourist APIs. If the user rated a suggestion as relevant, then the positive profile represents all positive reviews of that suggestion. The negative profile represents all negative reviews of the suggestion rated as irrelevant to the user. The aforementioned approaches consider different ranking features based on the similarity between candidate suggestions and positive and negative profiles. On the other hand, a learning to rank model exploiting 64 features using information obtained from Foursquare is presented by [83]. They used four groups of features: a) city-dependent features which describe the context (city) such as total number of venues in the city and total number of likes, b) category-dependent features that consist of the count of the 10 highest level categories obtained from Foursquare, c) venue-dependent features which describe the popularity of the venue in the city, and d) user-dependent features describing the similarity between user profiles and the suggestions. The most effective features were the venue-dependent features, that is, those indicating venue importance.

Besides recommendation, a critical part of our work is how to build test collections and create sub-collections from them. Because of this, we now introduce the topic and survey some of the most relevant works on that area. Creating a test collection is the most common approach for evaluating different Information Retrieval systems. Any test collection consists of a set of topics, a set of relevance assessments, and a set of retrievable documents. Since the beginning of IR evaluation by means of test collections, many researchers have looked at test collections from different angles. For example, what is the optimal number of topics to obtain reliable evaluations? In [156] the authors find that to have a reliable order of the systems, at least 50 topics have to be used in the evaluation stage. The problem of

analyzing the impact of different sub-collections (as a set of test collections) is also studied in the literature. In [148], the authors split TREC ad-hoc collections into two sub-collections and compared the effectiveness ranking of retrieval systems on each of them. They obtained a low correlation between the two rank runs, each run based on one of the two sub-collections. Later, in [145] a more exhaustive analysis is presented. The authors studied the impact of different sub-collections on the retrieval effectiveness by analyzing the effect over many test collections divided using different splitting approaches. Their study was based on runs submitted to two different TREC tracks, the `ad hoc` track from 2002 to 2008 and the `terabyte` one from 2004 to 2008. The authors found that the effect of these sub-collections is substantial, even affecting the relative performance of retrieval systems. In [146], the authors analyze the impact of the first-tier documents from ClueWeb09 collection in the effectiveness. The analysis was carried out on the TREC 2009 Web track, where participating teams were encouraged to submit runs based on Category A, and Category B. These categories were extracted from ClueWeb09 collection. Category A consists of 500 million English documents, Category B is a subset from Category A, it consists of 50 million documents of high quality seed documents and Wikipedia documents (they represent the first-tier documents). By analyzing the number of documents per subset and the relevance assessment, the authors found a bias towards Category B documents, in terms of assessed documents and those judged as relevant. In order to investigate this bias, they analyze the effect of first-tier documents on the effectiveness of runs based on Category A. First, they found that there is a high correlation between effectiveness and number of documents retrieved from the first-tier subset. Second, by removing all documents not from the first-tier subset, the effectiveness of almost all runs based on Category A was improved.

In the context of the CS track these questions arise again, since in this track participants share the same topics (profile, context) but they have to return a ranked list of documents for each topic, where these candidate documents can be selected from either the *Open Web* or *ClueWeb12* collection. Considering the potential impact that different collections may have on the retrieval effectiveness, one of our main interests in the rest of the chapter is to study the gap in effectiveness between *Open Web* systems and *ClueWeb12* systems in order to achieve reproducible results on a representative sample of the Web from *ClueWeb12* collection.

7.3 EXPERIMENTAL SETUP

7.3.1 *DataSet*

Our analyses are based on data collected from the TREC 2013 and 2014 Contextual Suggestion tracks (`CS 2013`, `CS 2014`). The CS track provides a set of profiles and a set of geographical contexts (cities in the United States) and the task is to provide a ranked list of suggestions (up to 50) for each topic (profile, context) pair. Each profile represents a single assessor past preferences for a given suggestion. Each user profile consists of two ratings per suggestion, on a 5-point scale; one rating for a suggestion's description as shown in the result list (i.e., a snippet), and another rating for its actual content (i.e., a web page). There are some differences between 2013 and 2014: First, the 50 target contexts used each year are not the same. Second, seeds cities from which the example suggestions were collected: in 2013 examples were collected from Philadelphia, PA, whereas in 2014 examples were collected from Chicago, IL and Santa Fe, NM. Third, the number of assessors also changed in these editions of the track. More details about the CS track can be found in the track's overview papers [81, 82], for 2013 and 2014, respectively.

The evaluation is performed as follows. For each topic – (profile, context) pairs – the top-5 documents of every submission are judged by the actual users whose profile is given (resulting in three ratings: description, actual document content, and geographical relevance assessments) and by NIST assessors (an additional rating for the geographical relevance assessment). Judgments are graded: subjective judgments range from 0 (strongly uninterested) to 4 (strongly interested) whereas objective judgments go from 0 (not geographically appropriate) to 2 (geographically appropriate). In both cases, a value of -2 indicates that the document could not be assessed (for example, the URL did not load in the judge's Web browser interface).

Documents are identified by their URLs (if they are submitted by runs based on *Open Web*) or by their *ClueWeb12* ids (if they are submitted by runs based on *ClueWeb12*). In our study, we use `ClueWeb12--qrels` to refer to relevance assessments of *ClueWeb12* documents, and `OpenWeb-qrels` to refer to relevance assessments of *Open Web* URLs, both sets of assessments built from the three relevance assessments files provided by the organizers: desc-doc-qrels, geo-user-qrels, and geo-nist-qrels.

The following metrics are used to evaluate the performance of the participating teams: Precision at 5 (P@5), Mean Reciprocal Rank (MRR), and a modified Time-Biased Gain (TBG) [80]. These metrics consider geographical and profile relevance (both in terms of document and description judgments), taking as thresholds a value of 1 and 3 (inclusive), respectively.

Table 49: Summary of judged documents form the *Open Web* and the *ClueWeb12* collection. The `total` column shows the total number of judged documents, while the `unique` presents the number of unique documents.

		total	unique	in *ClueWeb12*
CS 2014	*Open Web* runs	35,697	8,442	1,892
	ClueWeb12 runs	8,909	2,674	all
CS 2013	*Open Web* runs	28,849	10,349	2,894
	ClueWeb12 runs	7,329	3,098	all

7.3.2 *URL Normalization*

A recurring pre-processing step to produce the various results reported in the thesis concerns the normalization of URLs. We have normalized URLs consistently by removing their `www`, `http://`, `https://` prefixes, as well as their trailing "forwarding slash" character `/`, if any. In the special case of the URL referencing an `index.html` Web page, the `index.html` string is stripped from the URL before the other normalizations are applied.

7.3.3 *Mapping Open Web qrels to* ClueWeb12

We identify documents that are included in *OpenWeb-qrels* and exist in *ClueWeb12* collection (these documents are subsets 1 and 2 in Figure 16). We achieve this by obtaining the URLs from the *OpenWeb-qrels*, then, we search for these URLs in the *ClueWeb12* collection. To check the matching between qrels URLs and *ClueWeb12* document URLs, both were normalized as described in Section 7.3.2. We shared this subset with the CS track community[1]. In Table 49 we summarize the statistics derived from the *Open Web* and *ClueWeb12* relevance assessments in 2013 and 2014. We observe that the qrels do contain duplicates, that are not necessarily assessed the same. The differences can be explained by the CS track evaluation setup, where the top-5 suggestions per topic provided by each submitted run were judged individually [81, 82].

We have separated these documents into two subsets: subsets 1 and 2 from Figure 16. First, the subset 1 represents documents that were judged as *Open Web* documents and that have a matching *ClueWeb12* document, however they do not exist in *ClueWeb12* relevance assessments; we refer to this subset as (*OpenWeb-qrels-urls-in-ClueWeb12*). We consider these documents as additional judgments that can be

1 https://sites.google.com/site/treccontext/trec-2014/open-web-to-clueweb12-mapping

Table 50: URLs obtained from *Open Web* runs.

	2014	2013
Total number of URls	15,339,209	35,949,067
Unique number of URLs	75,719	102,649
Found in *ClueWeb12*	10,014	26,248

used to expand the *ClueWeb12* relevance assessments. The second subset consists of documents that overlap between *Open Web* and *ClueWeb12* relevance assessments – that is, they were judged twice –, we refer to this subset as *ClueWeb12-qrels* (*qrels-overlap*).

7.3.4 *Expanding ClueWeb12 qrels*

We expand the *ClueWeb12* relevance assessments by modifying the provided qrels files mentioned in Section 3.2.1. We achieve this by replacing in the qrels the URLs with their *ClueWeb12* ids (if they exist) based on the subset identified in Section 7.3.3.

7.3.5 *Mapping Open Web URLs to the* ClueWeb12 *documents Ids*

In this section, we describe how we map all URLs found by *Open Web* systems (in the submitted runs) to their *ClueWeb12* ids. We need this mapping to evaluate *Open Web* systems on *ClueWeb12* collection. In order to achieve this, we obtain the URLs from the *Open Web* runs. Then, we search for these URLs in *ClueWeb12* collection by matching the normalized URLs against documents normalized URLs in *ClueWeb12* collection. The result of this process is a mapping between URLs in the *Open Web* runs and their corresponding *ClueWeb12* ids (*OpenWeb-runs-urls-in-ClueWeb12*). Table 50 presents a summary about the *Open Web* URLs and the number of URLs found in *ClueWeb12* collection. As we see in the table, for CS 2013 around 25.6% of URLs have a matching document in *ClueWeb12*, while for CS 2014 only 13.2% exist in *ClueWeb12* collection.

7.4 COMPARING OPEN WEB AND CLOSED WEB RELEVANCE ASSESSMENTS

In this section we present an analysis to compare *Open Web* and *ClueWeb12* relevance assessments. In [49], we already showed that *Open Web* runs tend to receive better judgments than *ClueWeb12* results, based on analyzing the CS 2013 results. We repeat here the same experiment in order to investigate whether such tendency is still present in the 2014 test collection. We first compare *Open Web* and *ClueWeb12*

in general (the distribution of relevance assessments of documents returned by *Open Web* systems vs. those documents returned by *ClueWeb12* systems). Next, we focus on the documents in the overlap of the relevance assessments between *Open Web* systems and *ClueWeb12* systems.

7.4.1 *Fair Comparison*

In this section, we study **RQ5.1** *Do relevance assessments of Open Web differ (significantly) from relevance assessments of ClueWeb12 documents?* We analyze the distribution of profile judgments of documents returned by *Open Web* and *ClueWeb12* runs. In our analysis, we leave out the user, context, and system variables, and compare the judgments given to documents from the *Open Web* against those from *ClueWeb12*. In Figure 17, we observe that the *Open Web* histogram is slightly skewed towards the positive, relevant judgments. Even though we are not interested in comparing the actual frequencies. This would not be fair, mainly because there were many more *Open Web* submissions than *ClueWeb12* ones. Specifically, in TREC CS 2013, 27 runs submitted URLs from the *Open Web*, and only 7 runs used *ClueWeb12* documents. However, it is still relevant to see the relative frequency of -2's or -1's (document could not load at assessing time), used in CS 2013 and CS 2014, respectively. 4's (strongly interested) in each dataset: this is an important difference which will impact the performance of the systems using *ClueWeb12* documents.

Figure 18 shows the same analyses based on 2014 test collection. In that year of the track, 25 runs submitted URLs from the *Open Web*, and only 6 runs used *ClueWeb12* documents. We find that the judgments of documents from *Open Web* are skewed towards the positive (relevant) side, while judgments of documents from *ClueWeb12* are – again – skewed towards the negative (not relevant) part of the rating scale, similar to the findings on the 2013 test collection.

7.4.2 *Comapring Identical Documents from Open Web &* ClueWeb12

In Section 7.3.3, we identified two subsets of overlap between *Open Web* and *ClueWeb12* results: first, *OpenWeb-qrels-urls-in-ClueWeb12* that maps URLs from *OpenWeb-qrels* to *ClueWeb12* collection, and *qrels-overlap* that contains documents that exist in both *OpenWeb--qrels* and *ClueWeb12-qrels*. Based on these datasets, we investigate RQ2: **RQ5.2** *Can we identify an overlap between Open Web systems and ClueWeb12 systems in terms of documents suggested by both?, how are those documents in the overlap judged?*

Figure 19 shows the distribution of relevance assessments of documents in *OpenWeb-qrels-urls-in-ClueWeb12* for both CS 2013 and CS 2014. We observe that the distribution of judgments of these docu-

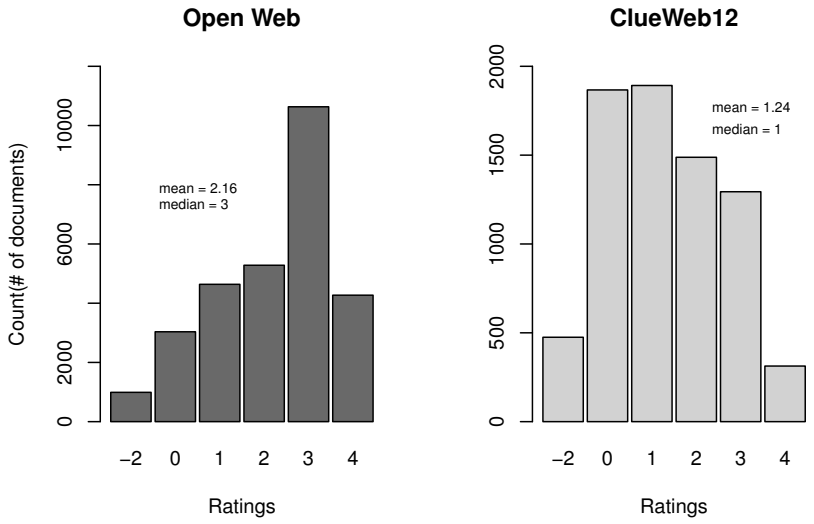


Figure 17: Judgments (document relevance) histogram of documents from *Open Web* (left) and from *ClueWeb12* (right). **CS 2013**

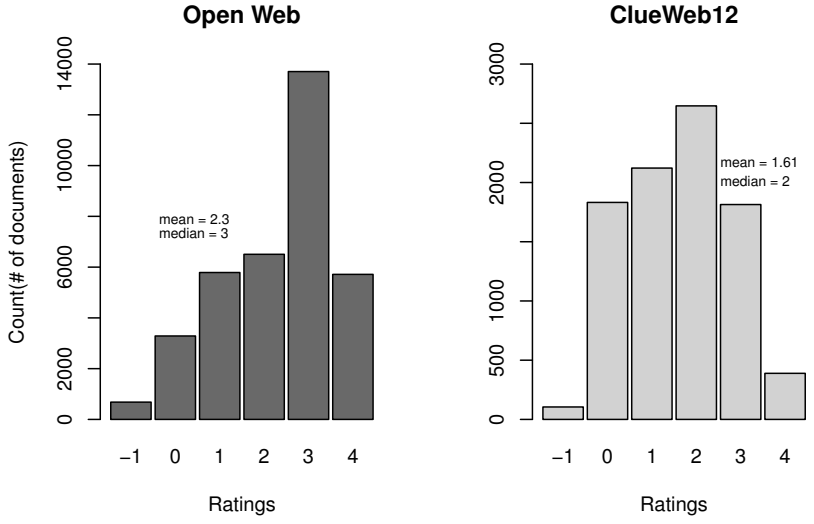


Figure 18: Judgments (document relevance) histogram of documents from *Open Web* runs (`left`) and *ClueWeb12* runs (`right`). **CS 2014**

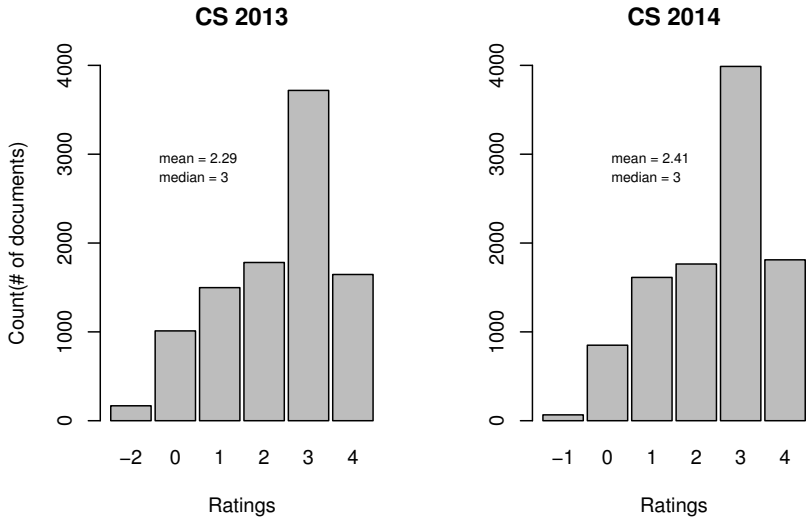


Figure 19: Judgments histogram of documents from *Open Web* qrels which exist in *ClueWeb12* collection for CS 2013 (left) and CS 2014 (right)

ments have a similar behavior as the whole *Open Web* judged documents. More precisely, we observe that the distribution is skewed towards the positive ratings when we look at 3 and 4 ratings for 2013 and 2014 datasets.

Now we focus on the *qrels-overlap* subset which contains documents shared by both *OpenWeb-qrels* and *ClueWeb12-qrels*. Our aim here is to detect any bias towards any of the document collections (the *Open Web* vs. *ClueWeb12*) based on the available sample of the judgments. In principle, the relevance judgments should be the same for the two sources, since in each situation the same document was retrieved by different systems for exactly the same user and context, the only difference being how the document was identified (as a URL or as a *ClueWeb12* id). Figure 20 and Figure 21 show how documents in the *qrels-overlap* were judged as *Open Web* URLs and as *ClueWeb12* documents in CS 2013 and CS 2014 test collections, respectively. We find that the documents in the overlap were judged differently. The judgments distributions of the documents shared by both *OpenWeb-qrels* and *ClueWeb12-qrels* suggest that there is a bias towards *OpenWeb--qrels* and this bias is consistent in 2013 and 2014 data. For CS 2013, part of the differences in judgments was attributed to a different rendering of the document for each source[2]. Assessors are influenced by several conditions, one of them is the visual aspect of the interface, but also the response time, the order of examination, the familiarity with the interface, etc. [114]. Therefore, it is important that these details are kept as stable as possible when different datasets are evaluated at the same time. It is also interesting to note that the number

2 Confirmed via email with the organisers for 2013 dataset.

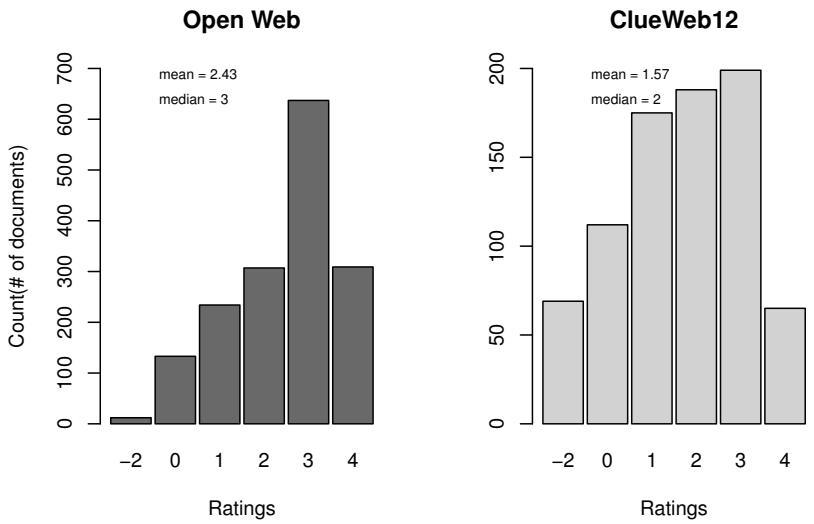


Figure 20: Judgments histogram of documents that exist in both *Open Web* qrels and *ClueWeb12* qrels. Figure on the (left) shows how these documents were judged as *Open Web* URLs, while the figure on the (right) shows how the same documents were judged as *ClueWeb12* documents. **CS 2013**

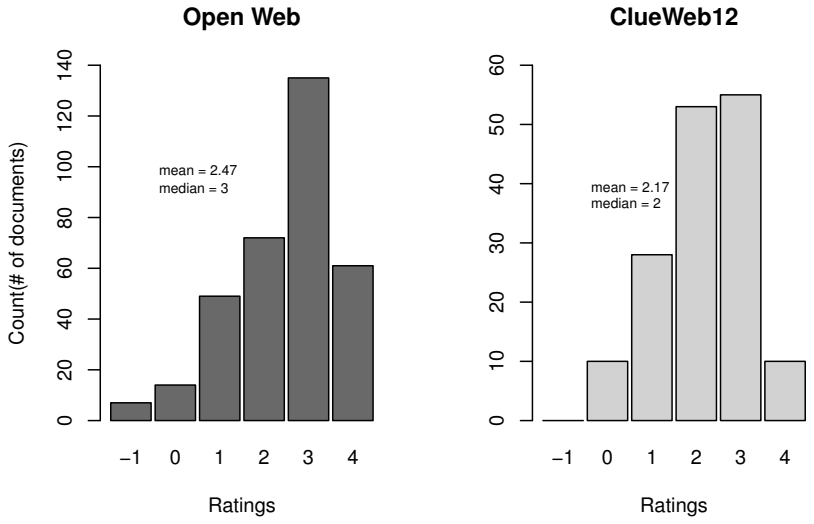


Figure 21: Judgments histogram of documents that exist in both *Open Web* qrels and *ClueWeb12* qrels. Figure on the (left) shows how these documents were judged as *Open Web* URLs, while the figure on the (right) shows how the same documents were judged as *ClueWeb12* documents. **CS 2014**

of *ClueWeb12* documents that could not load is higher in CS 2013 (−2) compared to CS 2014 (−1), probably due to the efforts of the organizers in the latter edition of running a fairer evaluation [78].

7.5 REPRODUCIBILITY OF *open web* SYSTEMS

In this section, we investigate **RQ5.3** *How many of the documents returned by Open Web systems can be found in the ClueWeb12 collection as a whole?* The goal of this analysis is to show how many of the results obtained by *Open Web* systems can be reproduced based on *ClueWeb12* collection. In Section 7.3.5, we presented the number of URLs found by *Open Web* systems and have a matching documents in *ClueWeb12* collection. Precisely in Table 50, we showed that for CS 2013 26,248 out of 102,649 URLs have a matching with *ClueWeb12* documents (25.6%), while for CS 2014 10,014 out of the 75,719 URLs (13.2%) have *ClueWeb12* documents match. In this section, we evaluate *Open Web* systems on *ClueWeb12* data. Analyzing the impact of *ClueWeb12* documents on the effectiveness of *Open Web* systems requires the following. First, we need to modify the *Open Web* runs using the *OpenWeb-runs-urls-in-ClueWeb12* dataset which has the mapping between *Open Web* URLs to *ClueWeb12* ids. Second – for evaluation completeness – we use the expanded *ClueWeb12-qrels* which was generated based on the *OpenWeb-qrels* URLs found in the *ClueWeb12* collection (*OpenWeb-qrels-urls-in-ClueWeb12* subset described in Section 7.3.4).

While modifying the *Open Web* runs, if the suggested URL has a matching in *ClueWeb12*, we replace the URL with its corresponding *ClueWeb12* id. If the URL has no match, then we skip the line containing that URL. We hence change the ranking after skipping those URLs. We present the effectiveness of original *Open Web* runs and the effectiveness of modified runs (replacing URLs with *ClueWeb12* ids), and we show the percentage of relative improvement in effectiveness of *Open Web* systems (on *Open Web* data vs *ClueWeb12*). Nonetheless, replacing the URLs with their matching *ClueWeb12* ids and pushing up their ranks by removing the URLs which have no *ClueWeb12* match will overestimate the performance and not show the corresponding impact on performance of those *ClueWeb12* documents if the ranking was preserved. To give an insight about the importance of *ClueWeb12* documents compared to the *Open Web* URLs that have no *ClueWeb12* match, we also include the percentage of *ClueWeb12* documents occurring in the top-5. To achieve this, when modifying the *Open Web* run, we replace the URLs with their match *ClueWeb12* ids, and keep the URLs as they are if they do not have a match. Then, for each topic, we compute the percentage of *ClueWeb12* documents in the top-5. The score for each run is the mean across all topics.

For CS 2013 systems (see Table 51) and for CS 2014 systems (see Table 52), we report the effectiveness of *Open Web* systems using their original run files as submitted to the track based on the original qrels (column named original). We report their effectiveness using the modified run files based on the expanded qrels as described above. Finally, we report the percentage of *ClueWeb12* documents in the top-5 as described above (how many *ClueWeb12* documents remain in the top-5 while preserving the URLs with no match).

In both tables, we observe the following: First, for some *Open Web* systems we were not able to reproduce their results based on *ClueWeb12* data, mainly because some systems have no matching at all with *ClueWeb12* collection. For systems that rely on the Yelp API to obtain candidate documents, we could not find any document whose host is Yelp in *ClueWeb12* collection, this is due to very strict indexing rules[3]. Second, we observe that the performance of *Open Web* systems decreases. However, this reduction in performance varies between systems, suggesting that pushing *ClueWeb12* documents up in the submitted rankings by removing URLs with no *ClueWeb12* id match has a different effect on each *Open Web* system. Third, some of top performing *Open Web* systems are performing very well when constrained to the *ClueWeb12* collection. For example, in the CS 2014 edition, UDInfoCS2014_2, BJUTa, and BJUTb systems even perform better than *ClueWeb12* systems (underlined systems in the table). Fourth, in terms of how representative *ClueWeb12* documents in the top-5, the percentage of *ClueWeb12* documents in the top-5 ranges from 1% to 46% (19% the mean across all *Open Web* systems, median=22%) for CS 2014 systems. For CS 2013, it ranges from 1% to 51% (22% the mean across all *Open Web* systems, median=25%)

3 See http://yelp.com/robots.txt

Table 51: Performance of *Open Web* systems on *Open Web* data vs. their performance on *ClueWeb12* data. Under each metric we present three values: original, replaced, and the relative improvement in effectiveness. The column named `original` presents the performance of submitted runs using the original qrels as provided by the organizers, whereas the column `replaced` shows the performance of modified runs (replacing URLs with their match *ClueWeb12* id and removing URLs with no match) using the expanded qrels. The % of *ClueWeb12* documents in top-5 column presents the percentage of *ClueWeb12* documents in the top-5 after replacing the URLs with their match *ClueWeb12* ids while preserving the ranks. The *ClueWeb12* systems (underlined) are included to show how they perform in comparison with *Open Web* systems evaluated on *ClueWeb12* data. For *ClueWeb12* systems no replacement has been applied, denoted by *n/a* under `replaced` and % of improvement. **CS 2013 systems**

	P@5			% *ClueWeb12*	MRR			TBG		
	original	replaced	%	in top-5	original	replaced	%	original	replaced	%
UDInfoCS1	0.5094	0.1444	-71.7	3.6	0.6320	0.2375	-62.4	2.4474	0.2273	-90.7
UDInfoCS2	0.4969	0.1379	-72.2	6.6	0.6300	0.2448	-61.1	2.4310	0.2993	-87.7
simpleScore	0.4332	0.1063	-75.5	3.1	0.5871	0.1974	-66.4	1.8374	0.1970	-89.3
complexScore	0.4152	0.1000	-75.9	3.5	0.5777	0.1500	-74.0	1.8226	0.1900	-89.6
DuTH_B	0.4090	0.1509	-63.1	24.9	0.5955	0.2999	-49.6	1.8508	0.4280	-76.9
1	0.3857	0.1688	-56.2	35.2	0.5588	0.3371	-39.7	1.5329	0.5450	-64.4
2	0.3731	0.1696	-54.5	32.6	0.5785	0.3144	-45.7	1.5843	0.5290	-66.6
udel_run_D	0.3659	0.1898	-48.1	39.8	0.5544	0.4182	-24.6	1.5243	0.7448	-51.1
isirun	0.3650	0.1568	-57.0	38.0	0.5165	0.2862	-44.6	1.6278	0.4265	-73.8
udel_run_SD	0.3354	0.1238	-63.1	25.4	0.5061	0.3131	-38.1	1.2882	0.4463	-65.4
york13cr2	0.3309	0.1198	-63.8	36.9	0.4637	0.2633	-43.2	1.3483	0.3762	-72.1
DuTH_A	0.3283	0.0991	-69.8	15.2	0.4836	0.2009	-58.5	1.3109	0.2287	-82.6
york13cr1	0.3274	0.1159	-64.6	36.9	0.4743	0.2667	-43.8	1.2970	0.3943	-69.6
UAmsTF30WU	0.3121	0.1182	-62.1	22.0	0.4803	0.2459	-48.8	1.1905	0.3626	-69.5
IRIT.OpenWeb	0.3112	0.1149	-63.1	25.0	0.4915	0.2492	-49.3	1.4638	0.4248	-71.0
CIRG_IRDISCOA	0.3013	0.1006	-66.6	23.0	0.4567	0.2010	-56.0	1.1681	0.2303	-80.3
CIRG_IRDISCOB	0.2906	0.1074	-63.0	24.3	0.4212	0.2042	-51.5	1.1183	0.2550	-77.2
uncsils_param	0.2780	no match	NaN	no match	0.4271	no match	NaN	1.3115	no match	NaN
uogTrCFP	0.2753	0.1000	-63.7	1.0	0.4327	0.3700	-14.5	1.3568	0.3784	-72.1
ming_1	0.2601	no match	NaN	no match	0.3816	no match	NaN	1.0495	no match	NaN
uncsils_base	0.2565	no match	NaN	no match	0.4136	no match	NaN	1.1374	no match	NaN
ming_2	0.2493	no match	NaN	no match	0.3473	no match	NaN	0.9673	no match	NaN
uogTrCFX	0.2332	0.0500	-78.6	0.8	0.4022	0.1562	-61.2	1.0894	0.1542	-85.8
run01	0.1650	0.1722	4.4	100.0	0.2994	0.3194	6.7	0.7359	0.7735	5.1
baselineB	0.1417	n/a	n/a	100.0	0.2452	n/a	n/a	0.4797	n/a	n/a
baselineA	0.1372	0.0841	-38.7	50.7	0.2316	0.1450	-37.4	0.5234	0.3001	-42.7
BOW_V17	0.1022	n/a	n/a	100.0	0.1877	n/a	n/a	0.3389	n/a	n/a
BOW_V18	0.1004	n/a	n/a	100.0	0.1971	n/a	n/a	0.3514	n/a	n/a
IRIT.ClueWeb	0.0798	n/a	n/a	100.0	0.1346	n/a	n/a	0.3279	n/a	n/a
RUN1	0.0628	n/a	n/a	100.0	0.1265	n/a	n/a	0.2069	n/a	n/a
csuio2	0.0565	no match	NaN	no match	0.1200	no match	NaN	0.1785	no match	NaN
csuio1	0.0565	no match	NaN	no match	0.1016	no match	NaN	0.1765	no match	NaN
RUN2	0.0565	n/a	n/a	100.0	0.1223	n/a	n/a	0.2020	n/a	n/a
IBCosTop1	0.0448	n/a	n/a	100.0	0.0569	n/a	n/a	0.1029	n/a	n/a

Table 52: Performance of *Open Web* systems on *Open Web* data vs. their performance on *ClueWeb12* data. Notation as in Table 51. **CS 2014 systems**

10.8

	P@5			% *ClueWeb12*	MRR			TBG		
	original	replaced	%	in top-5	original	replaced	%	original	replaced	%
UDInfoCS2014_2	0.5585	0.2275	-59.3	22.0	0.7482	0.5506	-26.4	2.7021	0.8604	-68.2
RAMARUN2	0.5017	no match	NaN	no match	0.6846	no match	NaN	2.3718	no match	NaN
BJUTa	0.5010	0.1781	-64.5	28.3	0.6677	0.3290	-50.7	2.2209	0.4752	-78.6
BJUTb	0.4983	0.1805	-63.8	29.5	0.6626	0.3319	-49.9	2.1949	0.4955	-77.4
uogTrBunSumF	0.4943	0.0769	-84.4	0.9	0.6704	0.1628	-75.7	2.1526	0.1690	-92.1
RUN1	0.4930	no match	NaN	no match	0.6646	no match	NaN	2.2866	no match	NaN
webis_1	0.4823	0.1768	-63.3	25.8	0.6479	0.3600	-44.4	2.1700	0.6195	-71.5
simpleScoreImp	0.4602	0.1283	-72.1	4.2	0.6408	0.2632	-58.9	1.9795	0.2595	-86.9
webis_2	0.4569	0.1768	-61.3	25.8	0.5980	0.3600	-39.8	2.1008	0.6195	-70.5
simpleScore	0.4538	0.1147	-74.7	5.4	0.6394	0.2368	-63.0	1.9804	0.2477	-87.5
run_FDwD	0.4348	0.1581	-63.6	30.6	0.5916	0.3390	-42.7	1.7684	0.5429	-69.3
waterlooB	0.4308	0.0932	-78.4	11.0	0.6244	0.2263	-63.8	1.8379	0.2686	-85.4
waterlooA	0.4167	0.0951	-77.2	12.0	0.6021	0.2280	-62.1	1.7364	0.2587	-85.1
UDInfoCS2014_1	0.4080	0.1278	-68.7	17.7	0.5559	0.2629	-52.7	1.6435	0.3185	-80.6
dixlticmu	0.3980	0.1735	-56.4	29.0	0.5366	0.3210	-40.2	1.5110	0.5240	-65.3
uogTrCsLtrF	0.3906	0.0667	-82.9	0.9	0.5185	0.0903	-82.6	1.9164	0.1285	-93.3
run_DwD	0.3177	0.1177	-63.0	25.8	0.3766	0.1718	-54.4	0.9684	0.1721	-82.2
tueNet	0.2261	0.0258	-88.6	2.6	0.3820	0.0452	-88.2	0.9224	0.0825	-91.1
choqrun	0.2254	0.1145	-49.2	33.2	0.3412	0.2223	-34.8	0.7372	0.3314	-55.0
tueRforest	0.2227	0.0258	-88.4	2.6	0.3604	0.0452	-87.5	0.9293	0.0825	-91.1
cat	0.2087	0.0954	-54.3	46.4	0.3496	0.1807	-48.3	0.6120	0.2544	-58.4
BUPT_PRIS_01	0.1452	0.1000	-31.1	16.2	0.4475	0.2982	-33.4	0.7453	0.3564	-52.2
CWI_CW12.MapWeb	0.1445	n/a	n/a	100.0	0.2307	n/a	n/a	0.6078	n/a	n/a
BUPT_PRIS_02	0.1425	0.0966	-32.2	17.4	0.3467	0.2080	-40.0	0.6601	0.2479	-62.4
gw1	0.1024	0.0386	-62.3	24.4	0.1694	0.0800	-52.8	0.3646	0.1150	-68.5
Model1	0.0903	n/a	n/a	100.0	0.1979	n/a	n/a	0.3411	n/a	n/a
lda	0.0843	0.0457	-45.8	30.4	0.1564	0.0928	-40.7	0.2461	0.1159	-52.9
Model0	0.0582	n/a	n/a	100.0	0.1023	n/a	n/a	0.1994	n/a	n/a
runA	0.0482	n/a	n/a	100.0	0.0856	n/a	n/a	0.1647	n/a	n/a
CWI_CW12_Full	0.0468	n/a	n/a	100.0	0.0767	n/a	n/a	0.1256	n/a	n/a
runB	0.0254	n/a	n/a	100.0	0.0552	n/a	n/a	0.0614	n/a	n/a

7.6 CONCLUSIONS

We have analyzed and discussed the balance between reproducibility and representativeness when building test collections. We have focused our analysis on the Contextual Suggestion TREC track, where in 2013 and 2014 it was possible to submit runs based on *Open Web* or based on *ClueWeb12*, a static version of the web. In both editions of the track, there were more runs based on *Open Web* compared to those based on *ClueWeb12* collection, which seems to go against any reproducibility criteria we may expect from such a competition. The main reason, as we have shown in this chapter, for that behavior is that systems based on *Open Web* perform better than systems based on *ClueWeb12* collection in terms of returning more relevant documents.

We have studied such difference in effectiveness from various perspectives. First, the analysis of relevance assessments of two years of the Contextual Suggestion track shows that documents returned by *Open Web* systems receive better ratings than documents returned by *ClueWeb12* systems. More specifically, we have found differences in judgment when looking at identical documents that were returned by both *Open Web* and *ClueWeb12* systems. Second, based on an expanded version of the relevance assessments – considering documents in the overlap of *Open Web* and *ClueWeb12* systems – and on generating *ClueWeb12*-based runs from *Open Web* runs, we have investigated the representativeness of *ClueWeb12* collection. Although the performance of *Open Web* systems decreases, we find a representative sample of *ClueWeb12* collection in *Open Web* runs.

8

IMPROVING CONTEXTUAL SUGGESTIONS USING OPEN WEB DOMAIN KNOWLEDGE

Contextual suggestion aims at recommending items to users given their current context, such as location-based tourist recommendations. Our contextual suggestion ranking model consists of two main components: selecting candidate suggestions and providing a ranked list of personalized suggestions. We focus on selecting appropriate suggestions from the *ClueWeb12* collection using tourist domain knowledge inferred from social sites and resources available on the public Web (*Open Web*). Specifically, we generate two candidate subsets retrieved from the *ClueWeb12* collection, one by filtering the content on mentions of the location context, and one by integrating domain knowledge derived from the *Open Web*. The impact of these candidate selection methods on contextual suggestion effectiveness is analyzed using the test collection constructed for the TREC Contextual Suggestion Track in 2014. Our main findings are that contextual suggestion performance on the subset created using *Open Web* domain knowledge is significantly better than using only geographical information. Second, using a prior probability estimated from domain knowledge leads to better suggestions and improves the performance.

8.1 INTRODUCTION

Recommender systems aim to help people find items of interest from a large pool of potentially interesting items. The users' preferences may change depending on their current context, such as the time of day, the device they use, or their location. Hence, those recommendations or suggestions should be tailored to the context of the user. Typically, recommender systems suggest a list of items based on users preferences. However, awareness of the importance of context as a third dimension beyond users and items has increased, for recommendation [32] and search [131] alike. The goal is to anticipate users' context without asking them. This problem – known as *contextual suggestion* in Information Retrieval (IR) and *context-aware recommendation* in the Recommender Systems (RS) community – is far from being solved. Depending on the type of context taken into account (time, location, group, short-term preferences, etc.), different techniques have been proposed. We use the definition of context stated in TREC's Contextual Suggestion (CS) track [79]: a context consists of a geographical location (a city and its corresponding state in the United States). The CS track investigates search techniques for

complex information needs that are highly dependent on context and user preferences. Submission based on documents collected from either the *Open Web* or *ClueWeb12* collection has been allowed since 2013, and the goal is to provide a list of ranked suggestions per (user, context) pair. An earlier analysis of the track's empirical results (in 2013 and 2014) has shown that runs based on the *Open Web* usually achieve higher effectiveness than those based on *ClueWeb12* collection [81, 82].

The majority of existing studies have relied on location-based social networks from the *Open Web* that are specialized in providing tourist suggestions, such as Yelp and Foursquare; focusing on re-ranking the candidate suggestions based on user preferences. The main problem addressed then is to model user interests through content-based recommendation, considering evidence in the form of terms taken from the textual descriptions [147] or categories [160] of suggestions in the user profile and their associated ratings, and approaches to rank suggestions based on their similarity with the user profile. Likewise, in [83] the authors combine various user-dependent and venue-dependent features, including the aforementioned descriptions and category features, in one ranking model. However, using the *ClueWeb12* collection as source of attractions requires first the selection of candidate documents, to be ranked later based on user preferences. The selection of candidate documents is a challenging task, since the (potentially) relevant suggestions have to be selected from this large collection.

In this chapter, we use domain knowledge inferred from location-based social networks on the *Open Web* for selecting suggestions from *ClueWeb12*. We evaluate our contextual suggestion model on two sub-collections of the *ClueWeb12* collection. One of the two sub-collections was generated using location-based social networks to annotate the candidate documents from *ClueWeb12* collection. We discuss how explicit representation of knowledge about the tourism domain available on the location-based social networks improves the effectiveness of our contextual suggestion model. We show that the same contextual suggestion model for recommendation achieves an order of magnitude difference in effectiveness, depending on the approach used to derive the candidate suggestions from *ClueWeb12*. We address the following research questions:

RQ6 *Can we identify a representative sample from the ClueWeb12 collection by applying filters from the Open Web tourist APIs tailored for the CS track?*

RQ6.1 *Do results differ based on the relevance dimensions considered (contextual vs profile relevance)?*

RQ6.2 *What is the impact of the type of domain knowledge inferred on recommendation effectiveness?*

RQ6.3 *Can we improve the results by modeling the candidate selection process probabilistically?*

8.2 EXPERIMENTAL SETUP

The models and approaches presented in this chapter have been evaluated by participating in the TREC 2014 Contextual Suggestion track (CS 2014). Our initial analysis is based on the two runs that our team submitted for evaluation. Both runs are based on sub-collections of candidate suggestions belonging to the *ClueWeb12* collection; the first using the **GeographicFiltered** sub-collection that we describe in Section 8.3.3.1 and the second one using the **TouristFiltered** sub-collection described in Section 8.3.3.2. In our analyses, we refer to these runs by the name of the sub-collection that it is based on.

8.3 CONTEXTUAL SUGGESTION MODEL

In this section, we formulate the problem and describe a general framework for finding and providing personalized recommendations based on user preferences. Then, we describe the two main components of our model. The first component represents our approach for generating personalized ranked suggestions to the user based on her preferences (Section 8.3.2). The second component describes our approach for modeling the selection of candidates from *ClueWeb12* collection (Section 8.3.3).

8.3.1 *General Model and Problem Formulation*

We assume that we have a set of suggestions – represented by a URL and a description – that have been judged by a set of users. The goal is to provide a ranked list of personalized suggestions for the users in new contexts. We exploit the user preferences and the given suggestion descriptions to model a textual user's positive and negative profiles into a similarity ranking model that is able to regulate the impact of the positive and negative profiles to generate a final scoring. We adopt a standard approach to content-based recommendation to determine a ranked list of suggestions:

$$P_{rel}(u, s) = P(s) \cdot SIM(u, s) \tag{8}$$

$P(s)$ is a probability that estimates how likely it is that suggestion s is relevant to the task, and controls the suggestions considered. We have experimented with different approaches to estimate this probability, described in detail in Section 8.3.3. Note that $P(s)$ does not necessarily depend on the user (the equivalent to the queries in traditional

retrieval models), although it may depend on the context; it can be compared to the "prior probability of relevance" of traditional information retrieval models. If the range of $P(s)$ is restricted to discrete values 0 and 1, then $P(s)$ acts as a Boolean filter that selects candidate suggestions based on some features.

8.3.2 *Personalization*

Similarity function $SIM(u, s)$ represents the (content-based) similarity between user interests and candidate suggestions, and determines the personalization of recommendations to the user's interests. We follow an approach to modeling user preferences that has been used widely in the literature on contextual suggestion; consider for example [144, 33, 147]. Descriptions of the previously rated attractions provide the basis to construct two user profiles for each user. The positive profile u^+ represents the attractions that the user u likes, whereas the negative profile u^- represents the attractions that the user u dislikes. We use the value 2.5 (since ratings are on 0 to 4 scale) as a threshold to discriminate between liked and disliked attractions. We compute the similarity score between a candidate suggestion s and a user u as follows:

$$SIM(u, s) = \lambda \cdot SIM(u^+, s) - (1 - \lambda) \cdot SIM(u^-, s) \tag{9}$$

where $SIM(u^+, s)$ is the similarity between user's positive profile and the candidate document, while $SIM(u^-, s)$ is the similarity between user's negative profile and the candidate document. λ is the parameter that regulates the contribution of the $SIM(u^+, s)$ and $SIM(u^-, s)$ to the final score. We used 5-fold cross-validation on training data to find the optimal $\lambda = 0.7$, which was selected from $[0, 1]$ in 0.1 steps. For this experiment, we considered the cosine similarity (based on term frequencies). This has been done after transforming the suggestions and the user profiles from text-representation into a weighted vector-based representation. In this transformation, we filter out the HTML tags from the content of the documents, apply common IR parsing techniques including stemming and stop-word removal.

8.3.3 *Selection Methods of Candidates*

The selection of candidate suggestions plays an important role for providing good suggestions to the users. We have already presented how previous works address the contextual suggestion challenge by using a variety of public tourist APIs – including Google Places, WikiTravel, Yelp, and Foursquare – to obtain a set of suggestions. Queries issued are usually related to the target context (location), either given by its name (i.e., *Chicago, IL*) or its latitude and longitude coordinates

(i.e., (41.85003, −87.65005)). Collecting suggestions from the *ClueWeb12* collection poses however new challenges, different from "just" constructing the right query to issue at location-based web services. We formulate the problem of candidate selection from *ClueWeb12* as follows. We have a set of contexts (locations) `C` – which correspond to US cities – provided by the `CS` track organizers. For each context $c \in C$, we generate a set of suggestions S_c from the *ClueWeb12* collection, which are expected to be located in that context. We investigate two different approaches toward generating S_c. The first approach is to apply a straightforward geographical filter, based on the content of the *ClueWeb12* documents. In the second approach, we exploit knowledge derived from external resources available on the *Open Web* about sites that provide touristic information, and apply this knowledge to *ClueWeb12* collection.

8.3.3.1 *Geographically Filtered Sub-collection*

Our main hypothesis in this approach is that a good suggestion (a venue) will contain its location correctly mentioned in its textual content. Therefore, we implemented a content-based geographical filter (named `geo_filter`) that selects documents mentioning a specific context with the format `(City, ST)`, ignoring those mentioning the city with different states or those matching multiple contexts. With this selection method we aim to ensure that the specific target context is mentioned in the filtered documents (hence, being geographically relevant documents). The documents that pass this filter form sub-collection, **GeographicFiltered**. In Equation (8), we express this geographic filtering process through probability $P(s)$, which defines the probability of a *ClueWeb12* document to be a candidate suggestion. In the simplest instantiation of our model, the probability of any document in *ClueWeb12* to be included in the **GeographicFiltered** sub-collection is assigned to 0 or 1 depending on whether it passes the geo_filter:

$$P(s) = \begin{cases} 1, & \text{if } (s) \text{ passes geo\_filter} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Approximately 9 million documents (8,883,068) from the *ClueWeb12* collection pass this filter.

8.3.3.2 *Applying Domain Knowledge to Sub-collection*

The sub-collection described in Section 8.3.3.1 only takes the context into account, however, users are not equally satisfied by any type of document when receiving contextual suggestions: they expect those documents to be *entertaining* [80]. This implies that documents about restaurants, museums, or zoos are more likely to be relevant than

stores or travel agencies [144]. We incorporate this information into our sub-collection creation process by sampling from the *ClueWeb12* collection considering knowledge from the tourist domain. In the following, we present alternative ways to select candidate documents from *ClueWeb12* collection using different filters. Each filter represents a domain knowledge about tourist information inferred from the *Open Web*.

DOMAIN-ORIENTED FILTER The first type of domain knowledge depends on a list of `hosts` that are well-known to provide tourist information, and are publicly available. We manually selected the hosts $\mathcal{H}$:= `{yelp, tripadvisor, wikitravel, zagat, xpedia, orbitz,` and `travel.yahoo}`. We consider these hosts as a domain filter to select suggestions from *ClueWeb12* collection. The probability of a document in *ClueWeb12* to be a candidate is either 0 or 1 depending only on its host. We define the probability $P(s)$ as:

$$P(s) = \begin{cases} 1, & \text{if host}(s) \in \mathcal{H} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

We refer to the set of documents that pass the domain filter defined in Equation (11) as *TouristSites*.

We assume pages about tourist information also have links to other interesting related pages, acknowledging the fact that pages on the same topic are connected to each other [75]. In order to maximize the extracted number of documents from the tourist domain we also consider the outlinks of documents from touristic sites. For each suggestion $s \in$*TouristSites*, we extract its outlinks `outlinks(s)` and combine all of them together in a set $\mathcal{O}$; including links between documents from two different hosts (`external links`) as well as links between pages from the same host (`internal links`). Notice that some of the outlinks may also be part of the *TouristSites* set, because of satisfying Equation (11). Next, we extract any document from *ClueWeb12* whose normalized URL matches one of the outlinks in $\mathcal{O}$. The probability of document s to be selected in this case is defined as:

$$P(s) = \begin{cases} 1, & \text{if URL}(s) \in \mathcal{O} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The set of candidate suggestions that pass this filter is called *TouristSitesOutlinks*.

ATTRACTION-ORIENTED FILTER We will now consider a different type of domain knowledge, by leveraging the information available

Table 53: Number of documents for each part of the **TouristFiltered** subcollection.

Filter	Number of documents
TouristSites	175,260
TouristSitesOutlinks	97,678
Attractions	102,604
TouristFiltered	375,542

on the Foursquare API [1]. For each context $c \in C$, we obtain a set of URLs by querying Foursquare API. If the document's URL is not returned by Foursquare, we use the combination of document name and context to issue a query to the Google search API e.g., "`Gannon University Erie, PA`" for name *Gannon University* and context *Erie, PA*. Extracting the hosts of the URLs obtained results in a set of 1,454 unique hosts. We then select all web pages in *ClueWeb12* from these hosts as the candidate suggestions, with its probability defined in the same way as in Equation 11. The set of documents that pass the host filter is referred to by *Attractions*.

Together, the three subsets of candidate suggestions *TouristSites*, *TouristSitesOutlinks* and *Attractions* form our second *ClueWeb12* subcollection that we refer to as **TouristFiltered**.

$$\textbf{TouristFiltered} := \mathit{TouristSites} \cup \mathit{TouristSitesOutlinks} \cup \mathit{Attractions}$$

Table 53 shows statistics about the documents that pass each filter.

8.3.3.3 *Candidates Selection Prior Probability*

In Sections 8.3.3.1 and 8.3.3.2, we introduced probabilities, used as binary filters so far, to decide which documents from the *ClueWeb12* collection should be selected as candidates. Each of these filters represents a different kind of knowledge related to tourism inferred from the *Open Web*. Now, we introduce three different methods to estimate prior $P(s)$ from the **TouristFiltered** sub-collection. Two non content-based priors exploit the correlation between relevance judgments, the depth of URLs, and the filters based on location-based social networks. The third prior is based on the content of the documents found by the best location-based filter. We evaluate the effect of these different estimations $P(s) = P^i_s$, where $i \in \{1, 2, 3\}$, by applying our contextual suggestion model on the **GeographicFiltered** sub-collection.

Previous research has shown that correlations between relevance and non content-based features such as document length can be exploited to improve retrieval results, e.g. [150]. Similarly, the authors of

1 `https://developer.foursquare.com/docs/venues/search`

Table 54: Distribution of *ClueWeb12* documents over URLs depth.

Depth	count	%
0	3,726,692	0.5
1	152,584,686	21.0
2	253,913,644	35.0
3	172,258,009	23.7
4	83,629,521	11.5
5	35,464,476	4.9
6	13,495,362	1.9
7	6,756,976	0.9
8	3,693,477	0.5
11	809,692	0.1

[120] presented a general model of embedding non content-based features of web pages (document length, in-link count, and URL depth) as a prior probability in the ranking model. By studying the correlation between the URL depth and the relevance of the webpage, they observed that the probability of being a home page is inversely related to URL depth. Motivated by these studies, we carry out a similar analysis on the URLs of *ClueWeb12* documents and the URLs of documents in the `CS` track ground truth. We use the number of slashes in the *normalized* URL to find the depth; a more fine-grained analysis like the four categories used in [120] is deferred to future work. Table 54 shows the depth distribution of URLs in the *ClueWeb12* collection. We estimate the relationship between URL depth and the prior probability of relevance by analyzing the ground truth of the *Open Web* qrels, the *ClueWeb12* qrels, as well as the URLs in the *Open Web* qrels that also exist in the *ClueWeb12* collection. We observe in Table 55 that approximately 72% of the documents in the *Open Web* qrels exist at the top levels of a website (depth zero and one), and that 75% of these are relevant, consistent with findings reported in the literature; we also find that the probability of a document being relevant is inversely related to the URL depth. However, the distribution of URL depth and their corresponding relevance is different for the *ClueWeb12* qrels, where the highest percentage of webpages presented (and relevant) in those runs are at depth two, one, and three (in that order).

We can now estimate a prior probability of relevance at each URL depth by combining the statistics derived from the qrels (based on the correlation between URL depth and relevance of the *ClueWeb12* ground truth information presented in Table 55 with the URL depth distribution of the complete collection, Table 54):

Table 55: Distribution of URLs depth over the documents in the *Open Web* qrels, and documents in the *ClueWeb12* qrels.

	Open Web runs					*ClueWeb12* runs			
	All		Relevant			All		Relevant	
depth	count	%	count	%	depth	count	%	count	%
0	23,657	66.31	9,271	67.69	0	159	1.79	22	2.53
1	2,113	5.92	636	4.64	1	1,856	20.89	208	23.88
2	6,957	19.50	2,758	20.14	2	4,537	51.06	479	54.99
3	2,211	6.20	853	6.23	3	1,412	15.89	86	9.87
4	434	1.22	113	0.82	4	688	7.74	57	6.54
5	179	0.50	47	0.34	5	168	1.89	13	1.49
6	52	0.15	5	0.04	6	43	0.48	3	0.34
7	61	0.17	6	0.04	7	9	0.10	1	0.11
8	14	0.04	8	0.06	10	9	0.10	2	0.23
11	1	0.00	13,697		13	4	0.05	871	
	35,679					8,885			

$$P_s^1 = P_s(\text{depth}) = P(\text{rel}|\text{URL}(\text{depth} = d_i)) = \frac{c(\text{Rel}, d_i)}{c(d_i)} \tag{13}$$

Similar to how we derive a prior probability of relevance from the URL depth data, we may also use the number of relevant documents generated by each subset filter to inform the prior probability of relevance. In this case, the probability of a document to be relevant considering that it has passed a filter is defined as follows:

$$P_s^2 = P_s(\text{filter}) = P(\text{rel}|\text{filter}_i) = \frac{c(\text{Rel}, \text{filter}_i)}{c(\text{filter}_i)} \tag{14}$$

Here, we use the statistics shown in Table 53 for the total number of documents that pass each **TouristFiltered** subset filter, to normalize the total number of relevant documents in each filter. The outcome is a filter-specific approach to estimate the prior probability of relevance. A document in **GeographicFiltered** sub-collection will get the prior probability of the filter that it passes, and the maximum prior is considered if multiple filters are satisfied. For the rest of the documents in **GeographicFiltered** sub-collection that do not satisfy any filter, they will get a prior estimated by the number of relevant documents in **GeographicFiltered** sub-collection normalized by its total number of documents.

The third prior P_s^3 is a content-based derived prior, where we use a language model constructed from documents that pass the best filter

Table 56: Distribution of URLs depth over the documents from *Open Web* qrels that exist in *ClueWeb12* collection.

	overlap			
	All		Relevant	
depth	count	%	count	%
0	8,847	87.78	1,891	81.54
1	473	4.69	180	7.76
2	423	4.20	149	6.43
3	210	2.08	52	2.24
4	78	0.77	19	0.82
5	36	0.36	17	0.73
6	11	0.11	3	0.13
7	1	0.01	0	0.00
8	13	0.13	8	0.34
	10,079		2,319	

in terms of highest performance values. Specifically, we learn from the documents that pass the *Attractions* filter which were part of the **TouristFiltered** run to compute the prior probabilities. The goal is to boost documents from **GeographicFiltered** sub-collection that are similar to the attraction documents. We construct two different language models. The first is from documents that pass the *Attractions* filter and were judged as relevant. The second is from documents that pass the *Attractions* filter and were judged as not relevant. After that, both sets are processed in a similar way to generate a language model: first the stop words and non-alphabetic words are removed; then, terms are ranked based on their relative frequency in each set.

8.4 SUB-COLLECTIONS DISCUSSION

In Chapter 7 we have shown that there exist documents in *ClueWeb12* that are relevant for the Contextual Suggestion task, namely because systems based on *Open Web* can still be competitive when the candidate documents are constrained to the *ClueWeb12* collection. However, the candidate selection process is very challenging, and the use of external, manually curated tourist services make this task easier, by promoting those relevant documents at the cost of reducing the reproducibility of the whole process.

In this section we aim to understand the candidate selection process and to provide recommendations in order to improve it. With this goal in mind, we study the **GeographicFiltered** and *Attractions* sub-collections by comparing the actual documents that pass the corresponding filters, so that we can analyze these sub-collections from

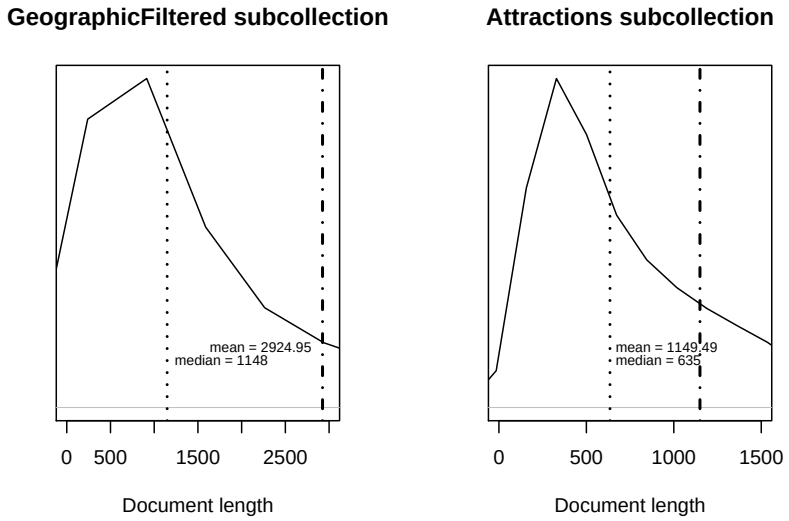


Figure 22: Distribution of the document length in words for the **GeographicFiltered** (left) and *Attractions* (right) sub-collections. Note the different range in the X axis.

the user perspective (what will the user receive?) instead of from the system perspective (what is the performance of the system?).

A first aspect we consider is the document length (in terms of words included in the processed HTML code), which gives an insight about how much information is contained (and shown to the user) in each sub-collection. We observe from Figure 22 that documents from the **GeographicFiltered** sub-collection are much larger than those from *Attractions*: their average length is twice as large as those from the other filter. This may suggest that relevant documents in the tourist domain should be short or, at least, they should not present too much information to the user. If this was true, it would be more interesting to retrieve – in the contextual suggestion scenario – home pages such as the main page of a museum or a restaurant, instead of their corresponding `Contact` or `How to access` sub-pages. Because of this, in the future we aim to take information about the URL depth into account when selecting the candidates, since it has been observed in [120] that the probability of being a home page is inversely related to its URL depth.

Related to the aforementioned aspect, we now want to check manually the content of some pages from each sub-collection. For this analysis we aggregate the judgments received to the documents submitted in each sub-collection, and then focus on documents with very bad or very good ratings in any of them. Specifically, we have found two candidate documents (presented in Figure 23) that clearly illustrates the main difference between these two sub-collections, and further corroborates the previous assumption: the **GeographicFiltered** subcollection requires pages where the target city and state

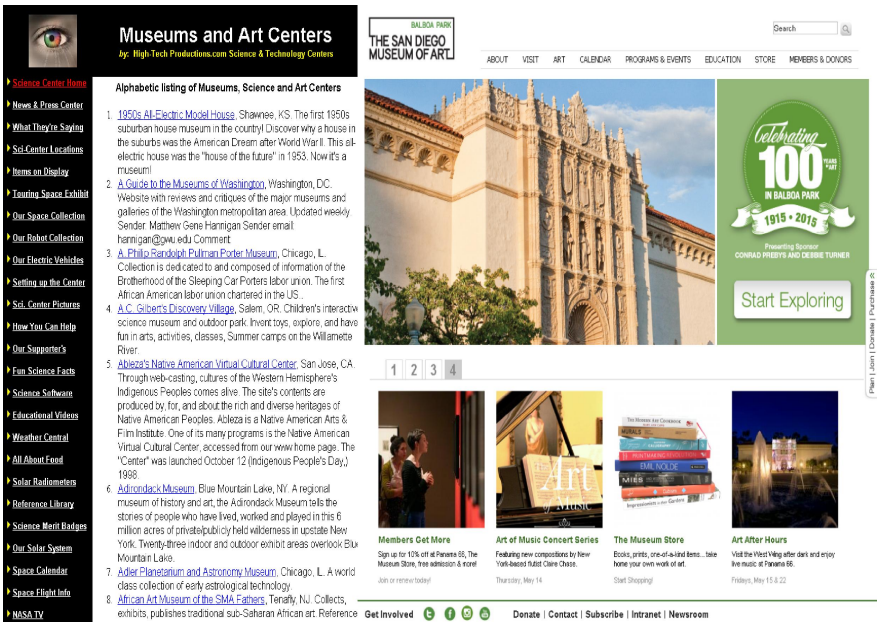


Figure 23: Screenshots of a document retrieved by the **GeographicFiltered** sub-collection (left) and by the *Attractions* sub-collection (right). The document in the left (clueweb12-0202wb-00-19744) was rated in average with a value of 1.9, whereas the one in the right (clueweb12-0200tw-67-19011) with a 3.

are present, which in turn favors pages containing listings of places located in that city, resulting in documents not very informative for an average tourist. On the other hand, the *Attractions* sub-collection tend to retrieve the home page of significant tourist places.

Finally, we have run an automatic classifier on the most popular terms used in each sub-collection in order to gain some insights about whether the content of the pages are actually different. We have used decision trees and decision rules and tried with different combinations of parameters (stemming, stopwords, confidence value for pruning, number of words to consider, etc.). Further experiments are needed to fully discriminate texts from each sub-collection, but some examples from our preliminary results show that the term *university* tends to appear more in documents from the **GeographicFiltered** sub-collection, whereas *park view* is more frequent in those in the *Attractions* sub-collection. In the future we want to exploit this information to improve the candidate selection process and the corresponding filters.

8.5 EFFECT OF USING EXTERNAL DOMAIN KNOWLEDGE FOR CANDIDATE SELECTION

In this section we study **RQ6** *Can we identify a representative sample from the ClueWeb12 collection by applying filters from the Open Web tourist APIs tailored for the CS track?* We compare the performance of our

contextual suggestion model (see Section 8.3) used to rank suggestions from the two presented sub-collections **GeographicFiltered** and **TouristFiltered**. We show empirically that the additional information acquired from location-based social networks provides the evidence needed to generate high quality contextual suggestions.

Table 57 summarizes the results from the evaluation, where we are initially only interested in the entries that take all relevance criteria into account, labeled by suffix `_all`. Clearly, the effectiveness using the **TouristFiltered** sub-collection outperforms the **GeographicFiltered** results by a large margin. Also, among the results obtained for the runs submitted in TREC 2014, the former approach was superior to all other submitted *ClueWeb12* runs, while the latter ranked near the bottom [82]. We should emphasize that the actual method that ranks the documents is exactly the same in both cases (Section 8.3.2), and hence, the difference in performance should be attributed to the differences in the candidate suggestions.

We can inspect the differences in more detail by comparing the two runs on a topic by topic basis. Table 58 shows the percentage of topics where the run based on the **TouristFiltered** sub-collection is better than, similar to, or worse than the run based on the **GeographicFiltered** sub-collection. We see that in approximately one third of the cases (it fluctuates per metric) the **TouristFiltered** sub-collection gives better results than **GeographicFiltered**, while for approximately 10% of the topics it leads to a lower performance. The fact that one method does not lead to improved results for every topic may indicate an opportunity to create a hybrid approach, where each topic is processed using the optimal sub-collection, following an approach based on query performance prediction, e.g., [63].

8.6 INSIGHTS ON THE RESULTS

We now present how we have addressed the three sub research questions mentioned at the beginning of the chapter and the results obtained in each situation. The measures are averaged after running a 5-fold cross-validation.

8.6.1 *Analysis per RelevanceDimensions*

In this section, we investigate **RQ6.1** *Do results differ based on the relevance dimensions considered (contextual vs profile relevance)?* Let us inspect the evaluation outcomes in more detail, by considering relevance dimensions individually. Recall that assessments are made considering geographical and profile relevance independently from each other. The latter one is further assessed as relevant based on the document or on the description provided by the method. Considering this information, we recomputed the evaluation metrics for each

Table 57: Performance of **GeographicFiltered** and **TouristFiltered** runs. Analysis per relevance dimension is considered; description (desc), document (doc), and geographical (geo) relevance. We denote with (all) when desc, doc, and geo relevance are considered.

Metric	**GeographicFiltered**	**TouristFiltered**
P@5_all	0.0431	0.1374
P@5_desc-doc	0.2081	0.2222
P@5_desc	0.2828	0.2788
P@5_doc	0.2620	0.2949
P@5_geo	0.1549	0.4808
MRR_all	0.0763	0.2305
MRR_desc-doc	0.2952	0.363
MRR_desc	0.394	0.4395
MRR_doc	0.3639	0.4718
MRR_geo	0.2166	0.6627
TBG	0.1234	0.5953
TBG_doc	0.1287	0.6379

Table 58: Comparison between the two runs based on **GeographicFiltered** and **TouristFiltered** sub-collections, by showing the percentage of topics where the **TouristFiltered** subcollection gives better, equal, or worse performance compared to the **GeographicFiltered** sub-collection.

TouristFiltered is	Better	Similar	Worse	Metric
than **GeographicFiltered**	33.11	58.53	8.36	P@5
	32.44	58.53	9.03	MRR
	41.47	47.49	11.04	TBG

topic while taking into account the geographical relevance provided by the assessors, as well as the description and document judgments, both separately and combined (that is, a document that is relevant both based on the description and when the assessor visited its URL).

Table 57 shows the effect of the relevance dimensions on the P@5 and MRR metric. When all the dimensions are considered (all), the **TouristFiltered** sub-collection is significantly better than the **GeographicFiltered** one. However, the difference in the performance between the two sub-collections decreases when we look at the relevance of a document and its description, that is, when we ignore the geographical aspect of the relevance. This means that both sub-collections are similar in terms of their appropriateness to the users. At the same time, we observe that the **TouristFiltered** sub-collection is more geographically appropriate, implying that using the domain knowledge to select the candidates improves the performance in that dimension. A similar observation is found when looking at the best relevance dimension, for the **GeographicFiltered** sub-collection it is the document description, whereas for the **TouristFiltered** sub-collection it is the geographical aspect, evidencing their pros and cons. The geographical, description, and document relevance assessments affect in the same way P@5 and the MRR metric.

8.6.2 *Impact of used Filters*

In this section, we investigate **RQ6.2** *What is the impact of the type of domain knowledge inferred on recommendation effectiveness?* We provide a deeper insight on why the domain knowledge-based sub-collection improves so much over the other sub-collection on the different relevance dimensions. Table 59 presents the contribution to the relevance dimensions of each of the **TouristFiltered** sub-collection subsets, where each subset was selected based on a different domain knowledge filter.

We start modifying the run based on the **TouristFiltered** sub-collection by computing effectiveness based only on suggestions from the *TouristSites* subset (second column), then we add to them suggestions from *TouristSitesOutlinks*, and finally suggestions from *Attractions* are added. The main conclusion drawn from this table is that the larger improvement in performance occurs after adding the candidates from *Attractions* subset. It is interesting to note that the performance of this part alone (last column) is comparable to that of the whole sub-collection.

8.6.3 *Effect of Prior Probability*

In this section, we investigate **RQ6.3** *Can we improve the results by modeling the candidate selection process probabilistically?* In this section, we

Table 59: Effect of domain knowledge filters on **TouristFiltered** run performance. Union means adding suggestions from the subset filter shown in column header of current column to the previous one. The percentage shows the relative improvement in effectiveness due to filter.

	TouristSites	∪ *TouristSitesOutlinks*		∪ *Attractions*		*Attractions*
Metrics	score	score	%	score	%	score
P@5_all	0.0392	0.0518	32.1	0.1374	165.3	0.1057
P@5_desc	0.0917	0.1200	30.9	0.2788	132.3	0.1973
P@5_doc	0.1008	0.1310	30.0	0.2949	125.1	0.2101
P@5_geo	0.2067	0.2659	28.6	0.4808	80.8	0.4667

Table 60: Effect of using a prior-probability of relevance on the **GeographicFiltered** run performance. *no prior* means applying the general ranking model with $P(s) = 1$ for documents that pass the geo_filter.

Metrics	no prior	depth prior	filter prior
P@5_all	0.0431	0.0660	0.1300
P@5_desc-doc	0.2081	0.1024	0.1912
P@5_desc	0.2828	0.1273	0.2350
P@5_doc	0.2620	0.1468	0.2579
P@5_geo	0.1549	0.3515	0.4842
TBG	0.1234	0.3007	0.5574
TBG_doc	0.1287	0.3281	0.5988

investigate the effect of adding a prior probability that we discussed in Section 8.3.3.3 on the performance of the contextual suggestion model. Table 60 shows the effect of depth prior, and the effect of the filter prior when applying the contextual suggestion model on the **GeographicFiltered** sub-collection. As shown in this table, there is a significant improvement on the performance of the **GeographicFiltered** sub-collection after applying the two priors independently. We observe that the domain filter prior has more impact on the performance.

Next, we study the effect of the third prior, which is a content-based derived prior, where we use a language model constructed from documents that pass the *Attractions* filter which were part of the **TouristFiltered** run. We experimented with different cut-offs for selecting the top words to form the language model, precisely the top 500, 1,000, and 5,000 words. Without finding a clear relation between cutoff and performance, we present results based on the top

1,000 terms. Table 61 shows the effect of using the similarity between the language models and the **GeographicFiltered** documents as prior. We observe that the performance is worse than without a prior (compare with first column of Table 60). However, this can also be explained by analyzing the number of documents that have judgments in the rankings generated by each method. We therefore reported also the percentage of judged documents in top-5 as well as the percentage of relevant documents among the judged, and the precision@5 with a condition that the document is judged. We now conclude that the language model generated from the relevant documents improves the performance.

Table 61: Language model constructed from relevant and not relevant documents.

Metrics	¬rel	rel
P@5_all	0.0034	0.0067
P@5_doc	0.0444	0.0694
%judged@5	28.55	46.73
%rel of judged@5	38.18	54.75
P@5_doc(judged)	0.2185	0.4824

8.7 CONCLUSION

We have presented an approach for improving contextual suggestions based on *ClueWeb12* collection. Our approach focused on selecting candidate documents from a large Web crawl (*ClueWeb12*), using tourist domain knowledge inferred from the location-based social networks from the *Open Web*. First, we presented Boolean filters for modeling selection of candidate suggestions, where each filter represents a different type of knowledge about the tourist domain. The filter is then integrated in the ranking model via a prior probability of relevance. Our empirical evaluation shows that using domain knowledge drawn from location-based social networks improves the performance of the contextual suggestion model when compared to the performance of the same ranking model, using the **GeographicFiltered** sub-collection that is created without any domain knowledge. Second, we found that the two sub-collections have different correlations with the dimensions of relevance considered in the evaluation (geographical and profile relevance), which opens up to investigate more the relation between the filters and the relevance dimension. Third, our analysis shows that filters used to create the **TouristFiltered** sub-collection vary in impact on contextual suggestion effectiveness. We exploit the knowledge of each filter to estimate a probability prior

embedded in the ranking model using 5-fold cross-validation analysis. We also consider the correlation between URL depth of the document and its relevance, as an alternative prior. The results of this analysis on the **GeographicFiltered** sub-collection suggest that both priors improved the performance. The domain filter prior has more influence on the performance, suggesting that the domain knowledge filter captures relevance better than the depth prior. In the future, we aim to investigate the effect of the filter prior by incorporating different sources of information, such as the relation between the filter criteria and URL depth, and the relation between filter criteria and the individual dimensions of relevance.

9 CONCLUSIONS

We presented our research of analyzing large-scale Web archives. In Part I of the thesis, we applied large-scale analysis of the content of different crawls / archives from the Web. In order to answer our research questions, we applied our analysis on different Web archives or crawls collected from the Web. In Chapter 3 and Chapter 4, we based our analysis on a part of the Dutch Web archive, which has been preserved by the National Library of the Netherlands (KB). In Chapter 5, we based our analysis on a part of the Dutch Web archive, and a crawl collected by Common Crawl.

We investigated what has been archived, and from that found traces to Web pages that were not archived (Chapter 3). Based on the link structure, we found that the number of unarchived pages is equal to the number of archived pages, i.e., only half of the target pages have been archived. We used the link *Anchor Text* to represent the unarchived pages. We showed that *Anchor Text* can be a useful resource to make unarchived pages retrievable among the archived pages; where archive pages have their raw content available in the archive. In this study, we showed how to expand the coverage of the Web archive by using aggregated *Anchor Text* as a representation of unarchived pages. Then, we showed how to use *Anchor Text* combined with timestamps to provide an estimation of what was popular on the Web at or before the crawling time (Chapter 4). Here, the analysis was based on a *depth-first* archive collection which uses a selection-based approach to select websites to be archived.

The content in an archive depends on the crawling strategy. For example, the websites included in the archive, the crawling depth and frequency. Therefore, we extended our analysis of using *Anchor Text* to estimate what was popular on the web by comparing *Anchor Text* of two Web crawls, each collected following a different crawling strategy (Chapter 5). Namely, a crawl from the Dutch Web archive collected using a *depth-first* strategy, and a crawl from *Common Crawl* collected using a *breadth-first* strategy

The presence of a Web page in an archive is highly dependent on the crawling process. Another source of bias that affects the accessibility of Web pages in the archive is the retrieval system used to find resources in the archive (Chapter 6). Retrievability has been used to quantify accessibility bias on community-collected collections such as TREC collections which are not real Web archives. The documents in Web archives are typically available in multiple versions which can be an implicit source of bias. We used the retrievability measure to

quantify the bias imposed by a retrieval system on documents in the Web archive collection.

In Part II of the thesis, we integrate knowledge that exists in the current Web (*Open Web*) to improve access to information from a crawled Web collection. We based our analysis on two years of participation in the Contextual Suggestion TREC track. The goal of this track is to provide personalized recommendation to users given their profile preferences and locations (*City*). Participating teams are allowed to suggest documents from the *Open Web* or from the *ClueWeb12* collection. In order to see if there is overlap between the two sets, and how the documents in the overlap were judged, we used the relevance assessments of both documents from the *Open Web* and those from the *ClueWeb12* collection. We expanded the test collection of the *ClueWeb12* collection using the documents from the *Open Web*. We found a large number of documents returned by *Open Web* systems that exist in the *ClueWeb12* collection, but were not retrieved by any of the *ClueWeb12* systems. We showed how to use them to expand the relevance assessment of the *ClueWeb12* collection (Chapter 7). Finally, we showed how to use the knowledge available in the *Open Web* to filter candidate documents from the *ClueWeb12* collection (Chapter 8).

9.1 MAIN FINDINGS

We summarize our main findings by answering the six main research questions that we posed in Chapter 1.

Using Link Structure & Anchor Text to Uncover and Reconstruct the Unarchived Web

Web archives preserve content of pages from the Web before they are lost. Despite the important attempts to preserve parts of the Web by archiving, a large part of the web's content remains unarchived. In practice it is not feasible to archive the entire web due to its ever-increasing size and rapidly changing content. The overall consequence is that our web archives are highly incomplete. In **RQ1**, we show how to increase the coverage of the archive by using the *Anchor Text*.

RQ1 *Can we uncover and provide representations of unarchived Web pages exploiting references to them from the archived Web pages?*

We tested our method on the Dutch Web archive that has been collected from the Web using a *depth-first* strategy based on a seed list of manually selected websites. We found that the archived pages contain evidence of a remarkable number of target pages and web sites that have not been archived (Section 3.3). The archive contains almost as many mentions of unarchived pages as the number of the

actually archived pages. While it is known that Web archives are incomplete, the assumption was the coverage of a depth-first crawl around the seed list to be higher.

Then, we build an implicit representation of missing unarchived pages based on link evidence and *Anchor Text* (Section 3.4). Anchor text is a short text which is used in the source page to describe the target page. In order to test the usefulness of the aggregated *Anchor Text* for providing description of the missing pages, we setup a known-item search experiment to investigate how it is easy to retrieve the unarchived pages based on the aggregation of their *Anchor Text* (Section 3.5). Anchor text is short text compared to the raw text content available for archived pages. Therefore, the main concern while testing the usefulness of *Anchor Text* was how they will be retrievable among other archived pages. We found that *Anchor Text* can be used to find and retrieve the page in the first ranks. The aggregated *Anchor Text* of unarchived pages has a skewed distribution, home pages have more unique words in their aggregated *Anchor Text* compared to the deeper pages (*non-homepages*). However, both unarchived homepages and non-homepages received similar satisfactory MRR average scores.

Our result shows that *Anchor Text* is useful to increase the effective coverage of the archive; while we have the content of archived pages, we can use *Anchor Text* to describe (a part of) the unarchived pages. This analysis is important for both the Web archive creators and the users of the Web archive, by being aware of the number of archived pages and what is missing from the archive but still can be reconstructed. For users searching the Web archive collection, it will be disappointing for them to get *missing page* or *unarchived page* message, knowing that the page existed in the Web, but was not archived. By adding the aggregated *Anchor Text* of unarchived pages, a partial recovery seems feasible and users will be able to get insight into the content of the missing pages. From the designer of the Web archive collection perspective, this analysis of *Anchor Text* and our analysis in the following two research questions may help in providing suggestion to expand the crawler's seeds list based on the link structure and *Anchor Text*.

There are some limitations in our study which can be extended in the future work. First, we applied our analysis on one year from the Dutch Web archive and our analysis was based on a one-year granularity. The analysis can be extended by considering a finer-grained time granularity, and on a longer time-frame. The Web archive collection consists of several crawls collected over time and the crawling frequency varies among websites in the seeds list. The content on the Web is dynamic, therefore, the aggregated *Anchor Text* of the unarchived page might change over time. Second, we applied our approach on a *breadth-first* archive collected using selection-based ap-

proach of choosing seeds list, which was expected to cover a high percentage of the target pages from the seeds list. We did not take into account the crawler settings, for example, excluded domains, and websites. Third, the approach of generating known-item queries. One way to automate this approach is to follow the query-simulation approach as in Chapter 6. Fourth, to measure the usefulness of aggregated *Anchor Text* to find unarchived page, we could build an index which has two representations of each page that has been archived; using the aggregated *Anchor Text*, and the actual raw text content.

In the following research question, we show how to use *Anchor Text* with their associated timestamps to reconstruct past popular topics. We use topic as in Section 4.1 and Section 5.1, to refer to user information needs which might consists of one or multiple words. One of the main advantages of using *Anchor Text* is that it is available in the Web archive for both the archived and the unarchived target pages. Thus, we still can get insight into what was popular on the Web even if some content is missing.

Using Anchor Text with Timestamps to Reconstruct Past Popular Topics

Queries that represent the past interests of real users, using the archived Web as it was, are usually not available, because they were not preserved. Initially the main purpose of Web archives creators was to preserve the Web (or at least part of it) before being lost, by repeatedly crawling the Web. Later on Web archive initiatives started to make their collection available for URL-based search. Only recently, some Web archive initiatives started to allow full-text search of their collection. Therefore, user-logs which represent a good resource of users implicit feedback for Web search, are not available for Web archive search. Motivated by studies which showed that *Anchor Text* is similar to documents titles and real users queries, and the lack of user query-logs, we propose to use the *Anchor Text* also to reconstruct the past popular topics, by making use of their associated timestamps.

RQ2 *Can we identify past popular topics using anchor text associated with hyperlinks of the Web archive?*

We used the link structure extracted from the Dutch Web archive to identify the most popular target hosts over time, and to get the most popular *Anchor Text* over time. The link structure was extracted from the *text/html* archived pages in the Dutch Web archive in the period between February 2009 and December 2012. First, we investigate the evolution of target hosts in the link structure (Section 4.3.1). We found that target hosts evolve significantly. Based on a one-month granularity, on average 25% among all hosts per month are new. After extracting, cleaning, de-duplicating and aggregating *Anchor Text*

(Section 4.2.2) and sorting them based on their frequency of use in the archive (Section 4.3.2), we need a way to evaluate them. For that we used the *WikiStats* dataset which has the aggregation of the number of views of Wikipedia pages over time (Section 4.2.3). By matching *Anchor Text* with titles of Wikipedia pages, we found that the matching is high for the most frequently used *Anchor Text*, and that matched *Anchor Text* covers Dutch entities, such as names of cities, newspapers (Section 4.3.3).

Based on this analysis, we conclude that *Anchor Text* can be used to help users exploring the archive and help them to understand what is in the archive. Finding the popular topic using *Anchor Text* can help choosing the crawling seeds.

In this study, the main limitation is the lack of user-logs. Queries over time that represent how users searched the Web archive collection are not available. Therefore, we looked into other sources for evaluating the representativeness of *Anchor Text* for past popular topics. We carried out our analysis on a Web archive collection created following a *depth-first* crawling strategy on a manually selected Web sites from the Dutch Web. In the following research question, we expand our analysis on a dataset collected from the entire Web following a *breadth-first* crawling strategy.

What is the Impact of the Crawling Strategy on Anchor Text Coverage of Past Popular Topics?

Web archives are created following a crawling strategy, thus the crawling strategy has a great influence on the data that is archived. In **RQ2**, we based our analysis on the *depth-first* dataset that has been crawled from the Dutch Web using a selection-based approach for selecting the websites to be archived. We extend our analysis of using temporal *Anchor Text* to identify past popular topics by studying two datasets collected from the Web following different crawling strategies. Precisely, the *depth-first* dataset is from the Dutch Web archive collected by National Library of the Netherlands (KB), and the *breadth-first* dataset is from Web crawls collected by the *Common Crawl* foundation.

RQ3 *How does the crawling strategy impact the Web archive's coverage of past popular topics?*

We explore how well the collections resulting from different crawling strategies cover content related to topics that were in the focus of Web users in a particular time period. We used *Anchor Text* from the link structure of the two collections crawled from the Web following two different crawling strategies. We had access only to the Web pages crawled in 2014 from *Common Crawl*. Therefore, we limited our analysis to pages crawled in 2014 in the collections. We used

three different sources that identify topics that were popular on the Web: Google Trends, *WikiStats*, and queries collected from users of the Dutch historic newspaper archive (Section 5.2.4). Our initial assumption was that the KB dataset would cover more topics from the Dutch domain, while the *Common Crawl* dataset would cover more global topics. To validate our assumption, for topics from Google trends and *WikiStats*, we split them into two groups; topics that were popular at global level and those that were popular in the Dutch domain. We found that the *breadth-first* dataset covers more topics, not only from the global but also topics from the *NL* domain. There are many differences between the *breadth-first* collection and the *depth-first* in terms of size, number of crawled Web pages and websites. Therefore, for a fair comparison, we generated subsets from the *Common Crawl* collection (Section 5.2.3), one subset based on the *.nl* domain, here we kept any link that belong to the *.nl*. The second subset based on the websites used by the KB to collect the Dutch archive, here, we kept any link that originates from a website in the KB seeds list. We found that the first subset covers more topics, and the second subset has comparable coverage compared to the KB collection, for both the global and the *NL* topics (Section 5.3).

Based on our analysis, we conclude that to increase the coverage of the archive of the popular topics the *breadth-first* strategy is preferable. We found that the *Common Crawl* collection (a *breadth-first* crawl) covers more topics than the KB collection (a *depth-first* crawl). This is not limited to popular topics from the entire Web but also applies to topics that were popular in the *.nl* domain. While it is difficult to answer the question which crawling strategy is better than the other, a mix of both would make sense based on the goal of the Web archive. For example, while applying a *depth-first* strategy of selected websites for completeness, the analysis of *Anchor Text* or other external resources should be used to identify additional pages that were popular. These may be then crawled at shallow level.

In our analysis to answer **RQ2** and **RQ3**, we used the same timeframe to compare *Anchor Text* and topics from different sources. While interesting topics might appear immediately on the live Web, it may take time until they are included in the archive. This might be addressed in future work.

In Chapter 5, we used different sources to identify past popular topics. There are some difference between these sources. For topics taken from the Google trends it might be clear that they are the most queries submitted by users and represent the most popular topics in the provided time-frame. The case is different for the *WikiStats* source, here pages' titles (topics) were treated the same by aggregating their views in the year on which we focused our analysis. In order to improve this, we could take into account the peak in page views across several time-frames. The third source that we used was

the queries submitted by users of the digitized new papers collection using Delpher Web service system. This collection contains digitized news papers articles published in period between 1618 and 1995, and the queries were submitted by users in 2015. Our motivation behind using these queries is that they are Dutch queries submitted by users to a Dutch collection, and we can view Web archive as a continuation of the digitized news papers archive. Also in the user-log of the Delpher system, there is a large number of queries with names and locations [152]. In the current Web, a high percentage of queries submitted to Web search engines consists of named entity queries [161, 134]. Motivated by this, N. Kanhabua et al. proposed an entity-oriented search system that supports retrieval from the Internet Archive [113].

Retrieval Bias Among Archived Web Documents

The Web content that can be made available to users depends on the crawling process. In different words, the accessibility of a Web page from the past depends on whether the page has been archived. In the previous research questions we focused on exploring the Web archive, in order to understand what has been archived, and what not. We used link structure and *Anchor Text* to uncover and provide representation of the unarchived pages. We showed the usefulness of this method to increase the coverage of the archive (**RQ1**). To compensate for the lack of user-logs, we used *Anchor Text* to estimate past popular topics (**RQ2**, and **RQ3**). Now, we shift our attention to the accessibility of the archived data. One way for users to access Web archives is through *full-text* search systems. Retrievability has been used in research to quantify accessibility bias on community-collected collections such as TREC collections. We explored the applicability of this approach on Web archives, where documents are typically available in multiple versions. This multiplicity of versions can be an implicit source of bias, that we quantify using this analysis.

The number of versions of each document varies depending on the frequency by which a specific website is crawled, and the point in time when the website was added to the crawler seeds. We explored how the retrieval bias is affected by this variance. We investigate whether search results in Web archives are influenced by varying number of versions, and how retrieval systems that are adapted to deal with them can be evaluated using retrievability. We investigate the suitability of retrievability for Web archive collection to measure the retrieval bias, and investigate how to rely on retrievability to evaluate systems that adapt and take into account the multiple versions.

RQ4 *What can we learn about Web archive access from studying the collection using a measure of retrievability?*

We used the retrievability measure to quantify retrieval bias induced by different retrieval systems on a subset of the Dutch Web

archive collection (from February 2009 until December 2012) from the National Library of The Netherlands[1] (KB). Here, the retrieval systems consider every version of a document in a Web archive as an independent document. We show that the retrievability of documents can vary for different versions of the same document, and that retrieval systems induce biases to different extents (section 6.5). Then we used retrievability to quantify the change in bias when the system is adapted to deal with multiple versions of a document (Section 6.6). We explored this using two approaches to collapse versions of the same document and thus refining the search results. First, we collapse document's versions based on their content similarity (*clustering*). Here, the cluster with more versions will get higher retrievability score. Second, we collapse the versions based on their *URL*. Here, we embedded a prior (based on the number of versions) with the scores given by retrieval systems, this means a document with more versions gets higher score. The clustering takes into account that the content of document's versions may change over time, and thus, collapse them into distinct clusters. The URL, considers them similar and collapse them into one (URL). The bias was lower for the two collapsing approaches, as compared when the systems which do not consider the multiple versions of the document. The three retrieval systems impose lower bias in the URL approach, as compared to clustering approach. Collapsing similar versions of the same document is a common practice in Web archive search systems, supported by our analysis where we showed the impact of the clustering behavior on the retrieval bias.

We investigate also whether the number of documents crawled in a particular year correlates with the number of documents in the search results from that year (Section 6.7). First, the analysis is based on documents' timestamps in the search results returned by the retrieval model for all queries, assuming queries are not inherently temporal in nature. The results show a relation between the number of documents per year and the number of documents retrieved by the retrieval system from that year. We further investigated the relation between the queries' timestamps and the documents' timestamps. First, we split the queries into different time-frames based on their timestamps using the one-year granularity. Then, we issued the queries against the retrieval model. The result show that our temporal queries indeed retrieve more documents from the assumed time-frame. Thus, the documents from the same time-frame were preferred by the retrieval model over documents from other time-frames. This means that the temporal queries increase the bias of documents from a specific time-frame, but this bias might be desirable. Issuing queries from a specific time-frame will help finding documents from that period and help exploring the content of the archive.

1 www.kb.nl

We have shown that retrievability is suitable for the assessment of Web archive retrieval systems, by showing its ability to capture the bias based on the approach followed to deal with multiple versions. This gives room for the designer of Web archive search systems to design their retrieval approach and rely on the retrievability measure for evaluation.

In order to compute the retrievability score of all documents in the collection, we need a set of queries to run against a given retrieval system. Ideally, we would use queries collected from users searching the collection. Again such a query log is not available for the Web archive. We used different approaches for generating the query set. First, we follow the approach used in [39] by simulating the queries from the content of the documents in the collection. Second, we use *Anchor Text* associated with hyperlinks in the Web archive. In our analysis of retrievability, we used the two query sets independently. We have left, the cross analysis between the two query sets for future work. The two query sets can be used to investigating the similarity between *Anchor Text* and the top single-term and bi-term queries drawn from the content of the document.

Integrating Online & Crawled Web

In the second part of the thesis, we investigated the integration of the *Open Web* and the archive Web in the context of the contextual suggestion. The specific nature of this track allows the participating teams to identify candidate documents either from the *Open Web* or from the *ClueWeb12* collection, a static version of the web. We compare the effectiveness of systems with a personalized algorithm based on top of documents collected from Web by relying on public tourist APIs, and systems that build their algorithms based on documents from the *ClueWeb12* dataset.

Reproducibility vs. Representativeness of Search Systems Built on Top of the Online (dynamic) Web and the Crawled (static) Web

RQ5 *Do relevance assessments of Open Web differ (significantly) from relevance assessments of ClueWeb12 documents? Can we identify an overlap between the two sets, and the documents in the overlap were judged?*

We focused our analysis on the Contextual Suggestion TREC track (*CS*), where in 2013 and 2014 it was possible to submit runs based on *Open Web* or based on *ClueWeb12*, a static version of the web. We based our analysis on the relevance assessments of documents from the *Open Web* and documents from the *ClueWeb12* collection. We found that documents returned by *Open Web* systems receive better ratings than documents returned by *ClueWeb12* systems. More specifically, we have found differences in judgment when looking at

identical documents that were returned by both *Open Web* and *ClueWeb12* systems (Section 7.4). Then, we looked at documents returned by *Open Web* systems that exist in the *ClueWeb12* collection, but have not been retrieved by the *ClueWeb12* systems. We used these documents to expand the relevance assessments of *ClueWeb12* systems. Based on an expanded version of the relevance assessments – considering documents in the overlap of *Open Web* and *ClueWeb12* systems – and on generating *ClueWeb12*-based runs from *Open Web* runs, we have investigated the representativeness of *ClueWeb12* collection (Section 7.5). Although the performance of *Open Web* systems decreases, we find a representative sample of the *ClueWeb12* collection in the *Open Web* runs.

Our analysis shows that result obtained from the *Open Web* needs a special care to make sure that results are reproducible. While content on the Web is very representative of real-time search, it is very dynamic, and hence there is risk that documents found by *Open Web* systems will be unavailable after some time. We contributed to the CS track by participating twice (2013 and 2014) in the track and raising the attention to the mapping between *Open Web* and *ClueWeb12* (see [2]). For the purpose of producing reproducible results, in the version of CS track in 2015, the organizers of the track introduce a pre-task for collecting candidate documents. At retrieval time, participating team use a fixed collection.

Using Knowledge Available in the Online Web to Annotate the Crawled Web

In Chapter 7 we have shown that there exist documents in *ClueWeb12* that are relevant for the Contextual Suggestion task, namely because systems based on *Open Web* can still be competitive when the candidate documents are constrained to the *ClueWeb12* collection. However, the candidate selection process is very challenging, and the use of external, manually curated tourist services make this task easier, by promoting those relevant documents at the cost of reducing the reproducibility of the whole process. We proposed an approach for selecting candidate documents from the *ClueWeb12* collection using the information available on the *Open Web*, and hence increasing the representativeness without scarifying the reproducibility.

RQ6 *Can we identify a representative sample from the ClueWeb12 collection by applying filters from the Open Web tourist APIs tailored for the CS track?*

Our contextual suggestion ranking model consists of two main components: selecting candidate suggestions from *ClueWeb12* collec-

2 `https://sites.google.com/site/treccontext/trec-2014/open-web-to-clueweb12-mapping`

tion and providing a ranked list of personalized suggestions (Section 8.3). To provide users with a good personalized list of documents, we need to first find the candidate documents that are geographically related. We focus on selecting appropriate suggestions from the *ClueWeb12* collection using tourist domain knowledge inferred from social sites and resources available on the public Web (*Open Web*). We compared the performance of two contextual suggestion (CS) systems which ranked suggestions from two sub-collections generated from the *ClueWeb12* collection (Section 8.3.3); the **GeographicFiltered**, and **TouristFiltered** generated using filters derived from the *Open Web*. Our empirical evaluation shows that using domain knowledge drawn from location-based social networks improves the performance of the contextual suggestion model when compared to the performance of the same ranking model, using the **GeographicFiltered** sub-collection that is created without any domain knowledge (Section 8.5).

Then, we investigated the evaluation of the two systems build on the two sub-collection in more details. The relevance assessments are made considering geographical and profile relevance independently from each other. The latter one is further assessed as relevant based on the document or on the description provided by the method. We found that the two CS systems build on top of the two sub-collections have different correlations with the dimensions of relevance considered in the evaluation (geographical and profile relevance), which opens up to investigate more the relation between the filters and the relevance dimension (Section 8.6.1).

After that, we investigated why the performance of our contextual suggestion model on the domain knowledge-based sub-collection (**TouristFiltered**) improves so much over the other sub-collection (**GeographicFiltered**) on the different relevance dimensions. Different filters were used to create the **TouristFiltered** sub-collection, where each filter represents a different type of knowledge about the tourist domain. The filter is then integrated in the ranking model via a prior probability of relevance (Section 8.6.2). Our empirical evaluation shows that using domain knowledge drawn from location-based social networks improves the performance of the contextual suggestion model when compared to the performance of the same ranking model, using the **GeographicFiltered** sub-collection that is created without any domain knowledge.

Finally, our analysis shows that filters used to create the **TouristFiltered** sub-collection vary in impact on contextual suggestion effectiveness. We exploit the knowledge of each filter to estimate a probability prior embedded in the ranking model using 5-fold cross-validation analysis. We also consider the correlation between URL depth of the document and its relevance, as an alternative prior (Section 8.6.3). The results of this analysis on the **GeographicFiltered** sub-

collection suggest that both priors improved the performance. The domain filter prior has more influence on the performance, suggesting that the domain knowledge filter captures relevance better than the depth prior. In the future, we aim to investigate the effect of the filter prior by incorporating different sources of information, such as the relation between the filter criteria and URL depth, and the relation between filter criteria and the individual dimensions of relevance. Our approach for selecting candidate documents from the *ClueWeb12* collection based on information obtained from the *Open Web* makes an improvement step towards partially bridging the gap in effectiveness between *Open Web* and *ClueWeb12* systems, while at the same time we achieve reproducible results on well-known representative sample of the web.

9.2 FUTURE WORK

The analysis that we carried out in Part I of the thesis was an exploratory study of the content in a large-scale Web archive.

Uncovering the Unarchived Web Study

In Chapter 3, we proposed an approach to use link structure and *Anchor Text* to uncover and provide representation of unarchived pages. This analysis was applied on a crawl which has been collected from the Web using a *depth-first* based on a manually selected websites. There are several possibilities for extending this study:

Apply on a breadth-first *Collection*
: This analysis can be applied on different Web crawls collected with different crawling strategy. For example, the *Common Crawl* collection which has been collected using a *breadth-first* strategy.

  The crawling strategy followed to create the archive has an influence on the content included in the archive. By following the *depth-first* crawling strategy, the target is to crawl as much as possible of pages from Web sites in the seeds list. By following the *breadth-first* crawling strategy, the goal is to discover as much as possible of the outgoing links but not in depth. Therefore, it is expected that the coverage of Web sites and top-level domains is higher. In a *breadth-first* dataset, the assumption is that the number and the diversity of Web pages linking to a target unarchived pages will be higher, and hence distinct *Anchor Text* will be used to link to the target page which might give different representation.

Along the Time-Axis
: In our analysis of uncovering and reconstructing missing pages

from the Web archive, we focused on one year of the Dutch archive. Another possibility for future work is to extend the analysis along the time axis. *How does the aggregated content of* Anchor Text *change over time for the missing pages?* Content on the Web is very dynamic, the aggregated *Anchor Text* might give different representation of the missing pages over time which might reflect on how the missing page change over time.

Consider Crawler's Settings

The Dutch web archive has been initiated in 2007, and the seed list keeps growing. Studying the impact of the crawler's seeds and the time when they have been included on the coverage of the crawl would be another direction of research. The settings of the crawler, such as the crawling frequency, inclusions and exclusion of websites, domains or top-level domains has an influence on the crawled data. Therefore, these settings will have an influence on the coverage of the Web archive. For example, the crawling frequency varies among websites in the seeds list. Considering the crawling frequency in a given time-frame while studying the archive's coverage was left for future work.

Reconstructing the Past Popular Topics

Our analysis of using *Anchor Text* to identify past popular topics in the Web was applied along the time axis using four years of a *depth-first* collection (Chapter 4). In Chapter 5, we extended this analysis on a *breadth-first* collection, we used a crawl from the *Common Crawl* collection. The goal was to compare the coverage of the past popular topics in two crawls collected with different crawling strategies. However, the analysis was limited to the crawl from 2014, mainly because we had access to this crawl which was hosted by *Surfsara*, a Dutch high performance computing center. In the following, we highlight possibilities for future work related to this study.

Along the Time Axis

All crawls collected by *Common Crawl* (since 2008 up to now) are available on Amazon as a public dataset. The availability of these crawls open the possibility for making the comparison of topics coverage along the time axis.

Inter-Crawl Analysis

Another possibility for future work, is to check the coverage of missing pages from the Dutch archive in the *Common Crawl* collection, especially the home pages and pages at shallow depth.

Topic Modeling of Anchor Text Along the Time Axis

In our analysis, we matched *Anchor Text* with topics taken from different sources to identify past popular topics using string

matching after applying basic preprocessing. This analysis can be extended by aggregating and modeling the topics taken from different into sources into broader topics, as well as *Anchor Text*.

Retrievability Bias

Based on our retrievability bias study in Chapter 6 there is room for future work.

Impact of Crawling Strategy
: We applied our analysis on a *depth-first* Web archive collected from manually selected Web sites. Therefore, most of the pages are all from the same Websites and are expected to be similar, which might influence the retrievability analysis starting from the queries drawn from the content of Web pages in the Web archive collection. This analysis can be extended by applying the retrievability on the *Common Crawl* collection which has been collected from the entire Web following the *breadth-first* crawling strategy, and thus the diversity of Web pages and content is expected to be higher.

Impact of Adding Representation of Unarchived Pages
: Expand the archive by adding the content of unarchived pages using our approach in Chapter 3 and investigate how the retrievability bias will be affected.

Modeling Queries Importance Over Time
: In our analysis, all queries drawn from the content of documents in the collection. We left modeling the weight of the queries based on their appearance over time for the future work.

Integrating Online & Crawled Web

Our research in Part II was about integrating the knowledge on the *Open Web* with crawled content. In the context of TREC Contextual Suggestion track, we analyzed the difference in effectiveness between systems using input documents from the online Web vs. system using input documents from crawled content from the Web (Chapter 7). In Chapter 8, we proposed an approach for selecting candidate documents from *ClueWeb12* collection using the information available on the *Open Web*. Our results are promising, and evidence that there is still room for improvement by using different and more information available on the *Open Web*. Content on the Web represent what is currently important or relevant for users of the Web. However, this information is very dynamic, content of Web pages changes and disappear overtime. Therefore, many Web archive initiatives started

crawling the Web (or at least part of it) repeatedly and keep it in Web archives. For tasks such as Contextual Suggestion which rely on sources from the Open web it is recommended to archive documents taken from the Web for the reusability purpose.

BIBLIOGRAPHY

[1] Dataset for learning to rank for wair research. https://code.google.com/p/pwa-technologies/wiki/L2R4WAIR.

[2] Amazon's alexa. URL http://www.alexa.com/, .

[3] Alexandria project. URL http://alexandria-project.eu/, .

[4] Internet archive archive-it service. URL https://archive-it.org/, .

[5] Archivespark – a github code repository. URL https://github.com/helgeho/ArchiveSpark, .

[6] National library of france. URL http://www.bnf.fr/en/tools/a.welcome_to_the_bnf.html.

[7] Common crawl. URL http://commoncrawl.org/.

[8] The largest and most comprehensive human-edited directory of the web. URL http://www.dmoz.org/.

[9] Apache hadoop. URL http://hadoop.apache.org/.

[10] Apache hbase. URL http://hbase.apache.org/.

[11] Heritrix. URL https://webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[12] International internet preservation consortium (iipc). URL http://www.netpreserve.org, .

[13] Tools and software setting up web archiving chain. URL http://www.netpreserve.org/web-archiving/tools-and-software, .

[14] Internet archive. URL https://archive.org/, .

[15] Internet world stats – usage and population statistics. URL http://www.internetworldstats.com/, .

[16] National library of the netherlands. URL http://www.kb.nl/.

[17] Longitudinal analytics of web archive data (lawa). URL http://www.lawa-project.eu/.

[18] Living web archives (liwa). URL http://liwa-project.eu/.

[19] Memento. URL http://mementoweb.org/about/, .

[20] Memento time travel. URL http://timetravel.mementoweb.org/, .

[21] List of web archives initiatives. URL http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[22] Apache pig. URL https://pig.apache.org/.

[23] Portuguese web archive – search interface. URL http://arquivo.pt/.

[24] The web robots pages. URL http://www.robotstxt.org/.

[25] Apache spark. URL http://spark.apache.org/.

[26] The british library. URL http://www.bl.uk/.

[27] Uk web archive. URL http://www.webarchive.org.uk/ukwa/.

[28] Warcbase – a github repository. URL https://github.com/lintool/warcbase, .

[29] Iso.2006. iso 28500: Information and documentation – the warc file format. URL http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf, .

[30] Internet archive wayback machine. URL https://archive.org/web/.

[31] *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, 2002. ACM.

[32] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, chapter 6, pages 217–253. Springer, Boston, MA, 2011. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3_7. URL http://dx.doi.org/10.1007/978-0-387-85820-3_7.

[33] M-Dyaa Albakour, Romain Deveaud, Craig Macdonald, and Iadh Ounis. Diversifying contextual suggestions from location-based social networks. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 125–134, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2976-7. doi: 10.1145/2637002.2637018. URL http://doi.acm.org/10.1145/2637002.2637018.

[34] James Allan, W. Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information

retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, 2012. doi: 10.1145/2215676.2215678. URL http://doi.acm.org/10.1145/2215676.2215678.

[35] Omar Alonso, Jannik Strötgen, Ricardo A Baeza-Yates, and Michael Gertz. Temporal information retrieval: Challenges and opportunities. *TWAW*, 11:1–8, 2011.

[36] Ahmed Alsum, Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel. Profiling web archive coverage for top-level domain and content language. In Trond Aalberg, Christos Papatheodorou, Milena Dobreva, Giannis Tsakonas, and Charles J. Farrugia, editors, *TPDL*, volume 8092 of *LNCS*, pages 60–71. Springer, 2013. ISBN 978-3-642-40500-6.

[37] Leif Azzopardi and Richard Bache. On the relationship between effectiveness and accessibility. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 889–890. ACM, 2010.

[38] Leif Azzopardi and Maarten de Rijke. Automatic construction of known-item finding test beds. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 603–604, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148276. URL http://doi.acm.org/10.1145/1148170.1148276.

[39] Leif Azzopardi and Vishwa Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 561–570, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458157. URL http://doi.acm.org/10.1145/1458082.1458157.

[40] Leif Azzopardi and Vishwa Vinay. Accessibility in information retrieval. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 482–489, 2008.

[41] Leif Azzopardi and Vishwa Vinay. Document accessibility: Evaluating the access afforded to a document by the retrieval system. In *Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pages 52–60. Citeseer, 2008.

[42] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. Building simulated queries for known-item topics: an analysis using six European languages. In *SIGIR 2007: Proceedings of the 30th*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, pages 455–462, 2007.

[43] Richard Bache and Leif Azzopardi. Improving access to large patent corpora. In *Transactions on large-scale data-and knowledge-centered systems II*, pages 103–121. Springer, 2010.

[44] Ricardo A. Baeza-Yates and Barbara Poblete. Evolution of the chilean web structure composition. In *LA-WEB*, pages 11–13, 2003.

[45] Shariq Bashir and Andreas Rauber. Analyzing document retrievability in patent retrieval settings. In *International Conference on Database and Expert Systems Applications*, pages 753–760. Springer, 2009.

[46] Shariq Bashir and Andreas Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1863–1866. ACM, 2009.

[47] Shariq Bashir and Andreas Rauber. Improving retrievability of patents in prior-art search. In *European Conference on Information Retrieval*, pages 457–470. Springer, 2010.

[48] Shariq Bashir and Andreas Rauber. On the relationship between query characteristics and IR functions retrieval bias. *Journal of the American Society for Information Science and Technology*, 62(8):1515–1532, 2011.

[49] Alejandro Bellogín, Thaer Samar, Arjen P. de Vries, and Alan Said. Challenges on combining open web and dataset evaluation results: The case of the contextual suggestion track. In de Rijke et al. [77], pages 430–436. ISBN 978-3-319-06027-9. doi: 10.1007/978-3-319-06028-6_37. URL `http://dx.doi.org/10.1007/978-3-319-06028-6_37`.

[50] Anat Ben-David and Hugo Huurdeman. Web archive search as research: Methodological and theoretical implications. *Alexandria*, 25(1-2):93–111, 2014.

[51] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A language modeling approach for temporal information needs. In *European Conference on Information Retrieval*, pages 13–25. Springer, 2010.

[52] Ilaria Bordino, Paolo Boldi, Debora Donato, Massimo Santini, and Sebastiano Vigna. Temporal evolution of the uk web. In *ICDM Workshops*, pages 909–918, 2008.

[53] Brian E. Brewington and George Cybenko. Keeping up with the changing web. *IEEE Computer*, 33(5):52–58, 2000.

[54] Andrei Z. Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[55] Adam Brokes, Libor Coufal, Zuzana Flashkova, Julien MasanÃšs, Johan Oomen, Radu Pop, Thomas Risse, and Hanneke Smu lders. Requirement analysis report "living web archive". URL `http://liwa-project.eu/images/uploads/d1-1.1_requirements_beg_v1.0.pdf`.

[56] Niels Brügger. Web history and the web as a historical source. *Zeithistorische Forschungen*, 9(2):316–325, 2012.

[57] Niels Brügger. Web historiography and internet studies: Challenges and perspectives. *New Media & Society*, page 1461444812462852, 2012.

[58] Niels Brügger. Historical network analysis of the web. *Social Science Computer Review*, 31(3):306–321, 2013.

[59] Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In Diana McCarthy and Shuly Wintner, editors, *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics, 2006. ISBN 1-932432-59-0. URL `http://acl.ldc.upenn.edu/E/E06/E06-1002.pdf`.

[60] Mike Burner and Brewster Kahle. Arc file format. URL `http://archive.org/web/researcher/ArcFileFormat.php`.

[61] Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, April 2001. ISSN 1046-8188. doi: 10.1145/382979.383040. URL `http://doi.acm.org/10.1145/382979.383040`.

[62] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15, 2015.

[63] David Carmel and Elad Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2010. doi: 10.2200/S00235ED1V01Y201004ICR015. URL `http://dx.doi.org/10.2200/S00235ED1V01Y201004ICR015`.

[64] CENTR. Domain wire stat report. Technical report, Council of European National Top Level Registrars (CENTR), 2013.

[65] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. 1999.

[66] Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. *ACM Trans. Internet Techn.*, 3(3):256–290, 2003.

[67] Miguel Costa and Mário J Silva. Understanding the information needs of web archive users. In *Proc. of the 10th International Web Archiving Workshop*, volume 9, page 6, 2010.

[68] Miguel Costa and Mário J. Silva. Characterizing search behavior in web archives. In *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011*, pages 33–40, 2011.

[69] Miguel Costa and Mário J. Silva. Evaluating web archive search systems. In Xiaoyang Sean Wang, Isabel F. Cruz, Alex Delis, and Guangyan Huang, editors, *WISE*, volume 7651 of *Lecture Notes in Computer Science*, pages 440–454. Springer, 2012. ISBN 978-3-642-35062-7.

[70] Miguel Costa, Francisco Couto, and Mário Silva. Learning temporal-dependent ranking models. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 757–766. ACM, 2014.

[71] Miguel Costa, Francisco M. Couto, and Mário J. Silva. Learning temporal-dependent ranking models. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *SIGIR*, pages 757–766. ACM, 2014. ISBN 978-1-4503-2257-7.

[72] Miguel Costa, Daniel Gomes, and Mário J Silva. The evolution of web archiving. *International Journal on Digital Libraries*, pages 1–15, 2016.

[73] Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *SIGIR*, pages 250–257. ACM, 2001.

[74] Na Dai and Brian D. Davison. Mining anchor text trends for retrieval. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan M. Rüger, and Keith van Rijsbergen, editors, *ECIR*, volume 5993 of *LNCS*, pages 127–139. Springer, 2010. ISBN 978-3-642-12274-3.

[75] Brian D. Davison. Topical locality in the web. In *In Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000*, pages 272–279. ACM Press, 2000.

[76] Michael Day. Preserving the fabric of our lives: A survey of web. In Traugott Koch and Ingeborg Solvberg, editors, *ECDL*, volume 2769 of *LNCS*, pages 461–472. Springer, 2003. ISBN 3-540-40726-X.

[77] Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors. *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, volume 8416 of *Lecture Notes in Computer Science*, 2014. Springer. ISBN 978-3-319-06027-9. doi: 10.1007/978-3-319-06028-6. URL http://dx.doi.org/10.1007/978-3-319-06028-6.

[78] Adriel Dean-Hall and Charles L. A. Clarke. The power of contextual suggestion. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, volume 9022 of *Lecture Notes in Computer Science*, pages 352–357, 2015. ISBN 978-3-319-16353-6. doi: 10.1007/978-3-319-16354-3_39. URL http://dx.doi.org/10.1007/978-3-319-16354-3_39.

[79] Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas, and Ellen M. Voorhees. Overview of the TREC 2012 contextual suggestion track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, volume Special Publication 500-298. National Institute of Standards and Technology (NIST), 2012. URL http://trec.nist.gov/pubs/trec21/papers/CONTEXTUAL12.overview.pdf.

[80] Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, and Paul Thomas. Evaluating contextual suggestion. In *Proceedings of the Fifth International Workshop on Evaluating Information Access (EVIA 2013)*, 2013.

[81] Adriel Dean-Hall, Charles L. A. Clarke, Nicole Simone, Jaap Kamps, Paul Thomas, and Ellen M. Voorhees. Overview of the TREC 2013 contextual suggestion track. In Ellen M. Voorhees, editor, *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, volume Special Publication 500-302. National Institute of Standards and Technology (NIST), 2013. URL http://trec.nist.gov/pubs/trec22/papers/CONTEXT.OVERVIEW.pdf.

[82] Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas, and Ellen M. Voorhees. Overview of the TREC 2014 contextual suggestion track. In Ellen M. Voorhees and Angela

Ellis, editors, *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, volume Special Publication 500-308. National Institute of Standards and Technology (NIST), 2014. URL http://trec.nist.gov/pubs/trec23/papers/overview-context.pdf.

[83] Romain Deveaud, M-Dyaa Albakour, Craig Macdonald, and Iadh Ounis. On the importance of venue-dependent features for learning to rank contextual suggestions. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1827–1830, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. doi: 10.1145/2661829.2661956. URL http://doi.acm.org/10.1145/2661829.2661956.

[84] Debora Donato, Stefano Leonardi, Stefano Millozzi, and Panayiotis Tsaparas. Mining the inner structure of the web graph. In *WebDB*, pages 145–150, 2005.

[85] Zhicheng Dou, Ruihua Song, Jian-Yun Nie, and Ji-Rong Wen. Using anchor texts with their hyperlink structure for web search. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *SIGIR*, pages 227–234. ACM, 2009. ISBN 978-1-60558-483-6.

[86] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR*, pages 459–460, 2003.

[87] Jonathan L. Elsas and Susan T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *WSDM*, pages 1–10, 2010.

[88] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th international conference on World Wide Web*, pages 669–678. ACM, 2003.

[89] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet L. Wiener. A large-scale study of the evolution of web pages. In *WWW*, pages 669–678, 2003.

[90] Atsushi Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In Jinpeng Huai, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins, and Xiaodong Zhang, editors, *WWW*, pages 337–346. ACM, 2008. ISBN 978-1-60558-085-2.

[91] Joseph L. Gastwirth. The estimation of the lorenz curve and gini index. In *The Review of Economics and Statistics. Vol. 54, No. 3*, pages 306–316, 1972. doi: 10.2307/1937992.

[92] Daniel Gomes and Mário J Silva. Characterizing a national community web. *ACM Transactions on Internet Technology (TOIT)*, 5 (3):508–531, 2005.

[93] Daniel Gomes and Mário J. Silva. Modelling information persistence on the web. In *ICWE*, pages 193–200, 2006.

[94] Daniel Gomes, André L. Santos, and Mário J. Silva. Managing duplicates in a web archive. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France, April 23-27, 2006*, pages 818–825, 2006.

[95] Daniel Gomes, André Nogueira, João Miranda, and Miguel Costa. Introducing the portuguese web archive initiative. In *8th international Web archiving workshop*. Springer, 2009.

[96] Daniel Gomes, João Miranda, and Miguel Costa. A survey on web archiving initiatives. In *TPDL*, pages 408–420, 2011.

[97] David Hawking and Nick Craswell. Very large scale retrieval and web search. In Ellen Voorhees and Donna Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005.

[98] David Hawking, Nick Craswell, Paul B. Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. *Computer Networks*, 31(11-16):1321–1330, 1999. doi: 10.1016/S1389-1286(99)00024-9. URL `http://dx.doi.org/10.1016/S1389-1286(99)00024-9`.

[99] David Hawking, Nick Craswell, Peter Bailey, and Kathleen Griffiths. Measuring search engine quality. *Inf. Retr.*, 4(1):33–59, 2001. doi: 10.1023/A:1011468107287. URL `http://dx.doi.org/10.1023/A:1011468107287`.

[100] Helen Hockx-Yu. The past issue of the web. In *Web Science 2011, WebSci '11, Koblenz, Germany - June 15 - 17, 2011*, pages 12:1–12:8, 2011. doi: 10.1145/2527031.2527050. URL `http://doi.acm.org/10.1145/2527031.2527050`.

[101] Helen Hockx-Yu. The past issue of the web. In *Web Science*, page 12. ACM, 2011.

[102] Helen Hockx-Yu. Access and scholarly use of web archives. *Alexandria: The Journal of National and International Library and Information Issues*, 25(1-2):113–127, 2014.

[103] Helge Holzmann, Vinay Goel, and Avishek Anand. Archivespark: Efficient web archive access, extraction and derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 83–92. ACM, 2016.

[104] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. The dawn of today's popular domains: A study of the archived german web over 18 years. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, pages 73–82, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4229-2. doi: 10.1145/2910896.2910901. URL http://doi.acm.org/10.1145/2910896.2910901.

[105] Gilles Hubert, Guillaume Cabanac, Karen Pinel-Sauvagnat, Damien Palacio, and Christian Sallaberry. IRIT, GeoComp, and LIUPPA at the TREC 2013 Contextual Suggestion Track. In *Proceedings of TREC*, 2013.

[106] HugoC. Huurdeman, Jaap Kamps, Thaer Samar, ArjenP. de Vries, Anat Ben-David, and RichardA. Rogers. Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries*, pages 1–19, 2015. ISSN 1432-5012. doi: 10.1007/s00799-015-0153-3. URL http://dx.doi.org/10.1007/s00799-015-0153-3.

[107] International Internet Preservation Consortium. Web Archiving Why Archive the Web? http://netpreserve.org/web-archiving/overview, 2014. Accessed: 2014-03-23.

[108] Rong Jin, Alexander G. Hauptmann, and ChengXiang Zhai. Title language model for information retrieval. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland* DBL [31], pages 42–48. doi: 10.1145/564376.564386. URL http://doi.acm.org/10.1145/564376.564386.

[109] Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3), 2007.

[110] Jaap Kamps. Web-centric language models. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *CIKM*, pages 307–308. ACM, 2005. ISBN 1-59593-140-6.

[111] Nattiya Kanhabua and Wolfgang Nejdl. On the value of temporal anchor texts in wikipedia. In *SIGIR 2014 Workshop on Temporal, Social and Spatially-aware Information Access (TAIA'2014)*, 2014.

[112] Nattiya Kanhabua, Roi Blanco, and Kjetil Nørvåg. Temporal information retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2015. ISSN 1554-0669. doi: 10.1561/1500000043. URL http://dx.doi.org/10.1561/1500000043.

[113] Nattiya Kanhabua, Philipp Kemkes, Wolfgang Nejdl, Tu Ngoc Nguyen, Felipe Reis, and Nam Khanh Tran. How to search the internet archive without indexing it. In *Research and Advanced Technology for Digital Libraries - 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings*, pages 147–160, 2016. doi: 10.1007/978-3-319-43997-6_12. URL http://dx.doi.org/10.1007/978-3-319-43997-6_12.

[114] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.

[115] Martin Klein and Michael L. Nelson. Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. *Int. J. on Digital Libraries*, 14(1-2):17–38, 2014.

[116] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[117] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999. ISSN 0004-5411. doi: 10.1145/324133.324140.

[118] Wallace Koehler. Web page change and persistence - A four-year longitudinal study. *JASIST*, 53(2):162–171, 2002. doi: 10.1002/asi.10018. URL http://dx.doi.org/10.1002/asi.10018.

[119] Marijn Koolen and Jaap Kamps. The importance of anchor text for ad hoc search revisited. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *SIGIR*, pages 122–129. ACM, 2010. ISBN 978-1-4503-0153-4.

[120] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland* DBL [31], pages 27–34. doi: 10.1145/564376.564383. URL http://doi.acm.org/10.1145/564376.564383.

[121] Reiner Kraft and Jason Zien. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA, 2004. ACM. ISBN 1-58113-844-X. doi: 10.1145/988672.988763.

[122] Page Lawrence, Brin Sergey, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[123] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1 (1), 2007. doi: 10.1145/1217299.1217301. URL http://doi.acm.org/10.1145/1217299.1217301.

[124] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *CIKM*, pages 469–475. ACM, 2003. ISBN 1-58113-723-0.

[125] Jimmy Lin, Milad Gholami, and Jinfeng Rao. Infrastructure for supporting exploration and discovery in web archives. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 851–856. ACM, 2014.

[126] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, chapter 3, pages 73–105. Springer, Boston, MA, 2011. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3_3. URL http://dx.doi.org/10.1007/978-0-387-85820-3_3.

[127] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 141–150, 2007.

[128] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

[129] Julien Masanès. *Web archiving*. Springer, 2006. ISBN 978-3-540-23338-1.

[130] Julien Masanès. *Web Archiving*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 3540233385.

[131] Massimo Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

[132] Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. Building enriched document representations using aggregated anchor text. In *SIGIR*, pages 219–226, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571981.

[133] Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer. Graph structure in the web - revisited: a trick of the heavy tail. In *WWW*, pages 427–432, 2014.

[134] Iris Miliaraki, Roi Blanco, and Mounia Lalmas. From "selena gomez" to "marlon brando": Understanding explorative entity search. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 765–775, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741284. URL https://doi.org/10.1145/2736277.2741284.

[135] Hannes Mühleisen. Wikistats – Wikipedia page views, 2013. URL http://wikistats.ins.cwi.nl.

[136] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web?: the evolution of the web from a search engine perspective. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *WWW*, pages 1–12. ACM, 2004. ISBN 1-58113-844-X.

[137] Christopher Olston and Marc Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

[138] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[139] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[140] M. Ras. Eerste fase webarchivering. Technical report, Koninklijke Bibliotheek, 2007.

[141] Marcel Ras and Sara van Bussel. Web archiving user survey. Technical report, Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), 2007.

[142] Andreas Rauber, Robert M. Bruckner, Andreas Aschenbrenner, Oliver Witvoet, and Max Kaiser. Uncovering information hidden in web archives: A glimpse at web analysis building on data warehouses. *D-Lib Magazine*, 8(12), 2002.

[143] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *CSCW*, pages 175–186, 1994.

[144] Andrei Rikitianskii, Morgan Harvey, and Fabio Crestani. A personalised recommendation system for context-aware suggestions. In de Rijke et al. [77], pages 63–74. ISBN 978-3-

319-06027-9. doi: 10.1007/978-3-319-06028-6_6. URL http://dx.doi.org/10.1007/978-3-319-06028-6_6.

[145] Mark Sanderson, Andrew Turpin, Ying Zhang, and Falk Scholer. Differences in effectiveness across sub-collections. In Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1965–1969. ACM, 2012. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398553. URL http://doi.acm.org/10.1145/2396761.2398553.

[146] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Effectiveness beyond the first crawl tier. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1937–1940. ACM, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063859. URL http://doi.acm.org/10.1145/2063576.2063859.

[147] Maya Sappelli, Suzan Verberne, and Wessel Kraaij. Recommending personalized touristic sights using google places. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 781–784, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484155. URL http://doi.acm.org/10.1145/2484028.2484155.

[148] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In Wei-Ying Ma, Jian-Yun Nie, Ricardo A. Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1063–1072. ACM, 2011. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010057. URL http://doi.acm.org/10.1145/2009916.2010057.

[149] M. Ángeles Serrano, Ana Gabriela Maguitman, Marián Boguñá, Santo Fortunato, and Alessandro Vespignani. Decoding the structure of the www: A comparative analysis of web crawls. *TWEB*, 1(2), 2007.

[150] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 21–29, New York, NY,

USA, 1996. ACM. ISBN 0-89791-792-8. doi: 10.1145/243199.243206. URL http://doi.acm.org/10.1145/243199.243206.

[151] Siddhi Soman, Arti Chharjta, Alexander Bonomo, and Andreas Paepcke. Arcspread for analyzing web archives. Technical report, Stanford InfoLab, 2012.

[152] Myriam C Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. Querylog-based assessment of retrievability bias in a large newspaper corpus. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 7–16. ACM, 2016.

[153] UNESCO. Charter on the preservation of digital heritage (article 3.4), 2003.

[154] H Van de Sompel, ML Nelson, and R Sanderson. RFC 7089 - HTTP framework for time-based access to resource states - Memento. RFC, Internet Engineering Task Force (IETF), 2013.

[155] Herbert Van de Sompel, Michael L Nelson, Robert Sanderson, Lyudmila L Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time travel for the web. *arXiv preprint arXiv:0911.1112*, 2009.

[156] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 316–323, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: 10.1145/564376.564432. URL http://doi.acm.org/10.1145/564376.564432.

[157] Stewart Whiting, Joemon M. Jose, and Omar Alonso. Wikipedia as a time machine. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 857–862, 2014. doi: 10.1145/2567948.2579048. URL http://doi.acm.org/10.1145/2567948.2579048.

[158] Colin Wilkie and Leif Azzopardi. Relating retrievability, performance and length. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 937–940. ACM, 2013.

[159] Peilin Yang and Hui Fang. An Exploration of Ranking-based Strategy for Contextual Suggestions. In *Proceedings of TREC*, 2012.

[160] Peilin Yang and Hui Fang. Opinion-based user profile modeling for contextual suggestions. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages

18:80–18:83, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2107-5. doi: 10.1145/2499178.2499191. URL http://doi.acm.org/10.1145/2499178.2499191.

[161] Xiaoxin Yin and Sarthak Shah. Building taxonomy of web search intents for name entity queries. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1001–1010, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772792. URL http://doi.acm.org/10.1145/1772690.1772792.

[162] Jonathan J. H. Zhu, Tao Meng, Zhengmao Xie, Geng Li, and Xiaoming Li. A teapot graph and its hierarchical structure of the chinese web. In *WWW*, pages 1133–1134, 2008.

Chapter 3 - Uncovering the Unarchived Web

This chapter presents how to use the link structure extracted from the crawled content to establish evidence of web pages outside the archive. They existed at crawling time, but were not archived. A reference or mention of the unarchived page in the content of the crawled pages included in the Web archive represents the *link evidence* of its existence. While going back in time to collect the uncrawled pages is impossible, we *reconstruct* basic representations of target URLs outside the archive which otherwise would have been lost.

Chapter 4 - Temporal Anchor Text as Proxy for Past User Queries

This chapter proposes an approach to reconstruct the information that would be provided by query log in the past using temporal Anchor Text. First, we study the link graph of four years of Web archive in order to show how the target hosts and Anchor Text evolve over time. Second, we investigate the importance of Anchor Text over time. Our approach is to rank Anchor Text based on popularity in the archive at specific points in time. Then, we check the importance of the top ranked Anchor Text in the public Web at the same time (time in the archive). In order to understand the importance of the Anchor Text, we rely on the *WikiStats* dataset, which provides an aggregation of page views of Wikipedia pages over time.

Chapter 5 - Comparing Topic Coverage in Breadth-first & Depth-first Crawls

This chapter considers the influence of the used crawling strategy on Anchor Text coverage of the past important topics. We perform our analysis on the links Anchor Text extracted from two Web crawls. One of our crawls was collected by the National Library of the Netherlands (KB) using a *depth-first* crawling strategy on manually selected websites from the *.nl* domain. The second crawl was collected by the *Common Crawl* foundation using a *breadth-first* crawling strategy on the entire Web. We used different sources as evidence of what the trending topics on the Web at the time of our used crawls. Since our crawls originate from the entire Web and from the Dutch domain, we looked for important topics both worldwide and on the national level. These sources are: Google Trends, *WikiStats*, and queries collected from users of the Dutch historic newspaper archive.

The two crawls differ in terms of size, number of crawled websites, and the domains of the crawled websites. Therefore, in order to allow fair comparison between the two crawls, we created sub-collections from the *Common Crawl* dataset based on the *.nl* domain and the KB seeds.

Chapter 6 - Quantifying Retrieval Bias in Web Archive Search

In the previous chapters, we used the link structure and Anchor Text, first to uncover and reconstruct the unarchived pages. One advantage of using anchor is its availability of both archived and unarchived target pages. Second, we used Anchor Text to reconstruct topics of interests to users in the past. In this chapter, we focus on the content of the crawled pages, pages that exist in the archive. One way to access Web archives is through full-text information retrieval systems. These retrieval systems influence which part of the Web archive is accessible, potentially imposing a retrieval bias among documents. This bias can be quantified using the *Retrievability* measure, a measure to quantify the relative likelihood of a document's retrieval over a large set of queries. In this chapter, we investigate the suitability of the retrievability measure to assess the retrieval bias induced by different retrieval systems among documents in a collection of *four years* of the Dutch Web archive.

Part II – Integrating Online & Crawled Web- Open Web (live and dynamic) & Crawled Web (archived and static)

In this we part, we consider the integration of the current web and the archived (crawled) web, the use case is the Contextual Suggestion TREC track.

Chapter 7 - The Strange Case of Reproducibility vs. Representativeness in Contextual Suggestion Test Collections

This chapter considers the task of integrating the current web and crawled web. The Contextual Suggestion (CS) TREC track provides an evaluation framework for systems that recommend items to users given their geographical context. The specific nature of this track allows the participating teams to identify candidate documents either from the *Open Web* or from the *ClueWeb12* collection, a static version of the web. We focus on analyzing reproducibility and representativeness of the *Open Web* and *ClueWeb12* systems. We study the gap in effectiveness between *Open Web* and *ClueWeb12* systems through analyzing the relevance assessments of documents returned by them. After that, we identify documents that overlap between *Open Web* and *ClueWeb12* results. We define two different sets of overlap: First, the overlap in the relevance assessments of documents

returned by *Open Web* and *ClueWeb12* systems, to investigate how these documents were judged according to the relevance assessments gathered when they were considered by *Open Web* or *ClueWeb12* systems. The second type of overlap is defined by the documents in the relevance assessments of the *Open Web* systems which are in *ClueWeb12* collection but not in the relevance assessments of *ClueWeb12* systems. The purpose is to use the judgments of these documents (mapped from *Open Web* on *ClueWeb12* collection) to expand the relevance assessments of *ClueWeb12* systems resulting on having a new test collection.

Chapter 8 - Improving Contextual Suggestions using Open Web Domain Knowledge

The majority of existing studies related to contextual suggestions have relied on location-based social networks from the *Open Web* that are specialized in providing tourist suggestions, such as Yelp and Foursquare; focusing on re-ranking the candidate suggestions based on user preferences. In this chapter, we consider the use of domain knowledge inferred from such location-based social networks on the *Open Web* for selecting suggestions from *ClueWeb12*.

SAMENVATTING

Het onderzoek in dit proefschrift focust zich op de analyse van omvangrijke webarchieven. In deel I van het proefschrift hebben we een grootschalige analyse gedaan van de inhoud van verschillende crawls / archieven van het web. Om onze onderzoeksvragen te beantwoorden hebben we onze analyse toegepast op verschillende webarchieven en crawls die zijn verzameld op het web. In hoofdstuk 3 en hoofdstuk 4 hebben we onze analyse gebaseerd op een deel van het Nederlands Webarchief, dat door de Koninklijke Bibliotheek (*KB*) wordt vastgelegd en bewaard. De analyse in hoofdstuk 5 is gebaseerd op een deel van het Nederlands webarchief en een crawl verzameld door Common Crawl. We zijn nagegaan wat er is gearchiveerd en in het gearchiveerde materiaal hebben wij sporen gevonden van webpagina's die niet zijn gearchiveerd (hoofdstuk 3). Op basis van de linkstructuur constateerden we dat het aantal niet-gearchiveerde pagina's gelijk is aan het aantal gearchiveerde pagina's. Dat wil zeggen dat slechts de helft van de pagina's waarnaar links verwijzen is gearchiveerd. We hebben de link *Anchor Text* gebruikt om de niet-gearchiveerde pagina's te representeren. We hebben aangetoond dat *Anchor Text* een nuttige bron kan zijn om niet-gearchiveerde pagina's terug te vinden onder de gearchiveerde pagina's, terwijl de inhoud van archiefpagina's beschikbaar is in het archief. In dit hoofdstuk hebben we aangetoond hoe de dekking van het webarchief uitgebreid kan worden door gebruik te maken van geaggregeerde *Anchor Text* als een voorstelling van niet-gearchiveerde pagina's. Vervolgens lieten we zien hoe *Anchor Text* in combinatie met tijdsaanduidingen gebruikt kan worden om een schatting te maken van wat op of vóór het tijdstip dat een pagina is vastgelegd populair was op het web (hoofdstuk 4). De analyse in dit hoofdstuk is gebaseerd op een *depth-first* archiefcollectie die een selectiegerichte aanpak gebruikt om websites te archiveren. De inhoud van een webarchief is afhankelijk van de crawlstrategie. Deze strategie omvat onder meer de websites die worden vastgelegd, de crawldiepte en de frequentie van archivering. Daarom hebben we onze analyse met het gebruik van *Anchor Text* om te beoordelen wat populair was op het web uitgebreid door het vergelijken van *Anchor Text* van twee webcrawls, die worden verzameld via verschillende crawlstrategieën (hoofdstuk 5). De eerste is een crawl uit het Nederlands Webarchief die is verzameld met behulp van een *depth-first* strategie, waar de tweede crawl van *Common Crawl* is, verzameld met behulp van een *breadth-first* strategie. De aanwezigheid van een webpagina in een archief is sterk afhankelijk van het crawlproces. Een andere bron van *bias* die de toegankelijkheid

van webpagina's in het archief beïnvloedt, is het retrievalsysteem dat gebruikt wordt om bronnen in het archief te vinden (hoofdstuk 6). Dit proefschrift gebruikt *retrievability* om de toegankelijkheids*bias* in gemeenschappelijk verzamelde collecties te kwantificeren, zoals TREC-collecties die geen echte webarchieven zijn. De documenten in webarchieven zijn meestal beschikbaar in meerdere versies die een impliciete bron van *bias* kunnen zijn. We hebben de *retrievability* measure gebruikt om de bias te kwantificeren die door een retrievalsysteem op documenten in de webarchiefcollectie wordt opgelegd. In deel II van het proefschrift integreren we kennis die bestaat in het huidige web (*Open Web*) om de toegang tot informatie uit een gecrawlde webcollectie te verbeteren. We hebben onze analyse gebaseerd op een tweejarige deelname aan de *Contextual Suggestion TREC track*. Het doel van deze track is om gebruikers persoonlijke aanbevelingen te doen op basis van hun profielvoorkeuren en locaties (*City*). Deelnemende teams mogen documenten van de *Open Web* of *ClueWeb12* collectie voorstellen. Om te kijken of er overlap tussen de twee sets bestaat en hoe de documenten in de overlap werden beoordeeld, hebben we de *relevance assessments* gebruikt van documenten, zowel in de *Open Web* als in de *ClueWeb12* collectie. We hebben de testcollectie van de *ClueWeb12* collectie uitgebreid met behulp van de documenten van de *Open Web* en vonden een groot aantal documenten dat door de *Open Web* systemen is teruggegeven en in de *ClueWeb12* collectie bestaat, maar niet door een van de *ClueWeb12* systemen werd opgehaald. We hebben laten zien hoe deze documenten kunnen worden gebruikt om de relevantiebeoordeling van de *ClueWeb12*-collectie uit te breiden (hoofdstuk 7). Tenslotte hebben we aangetoond hoe de beschikbare kennis in de *Open Web*-collectie toegepast kan worden om kandidaatdocumenten uit de collectie *ClueWeb12* te filteren (hoofdstuk 8).

LIST OF PUBLICATIONS

Part I – Studying Large-Scale Web Archives- Accessibility of Web Archive Content Along the Time Axis

1. *Uncovering the unarchived web.* **Thaer Samar** and Hugo C. Huurdeman and Anat Ben-David and Jaap Kamps and Arjen P. de Vries. The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014.

2. *Finding pages on the unarchived Web.* Hugo C. Huurdeman and Anat Ben-David and Jaap Kamps and **Thaer Samar** and Arjen P. de Vries. IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014.

3. *Lost but not forgotten: finding pages on the unarchived web.* Hugo C. Huurdeman and Jaap Kamps and **Thaer Samar** and Arjen P. de Vries and Anat Ben-David and Richard A. Rogers. International Journal on Digital Libraries, 2015.

4. *Temporal Anchor Text as Proxy for Real User Queries.* **Thaer Samar** and Arjen P. de Vries. Proceedings of the 5th International Workshop on Semantic Digital Archives (SDA 2015) co-located with the International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015.

5. *Comparing Topic Coverage in Breadth-first & Depth-first Crawls using Anchor Texts*. Published at TPDL'16 by **Thaer Samar**, Myriam C. Traub, Jacco van Ossenbruggen, and Arjen P. de Vries. Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 9-9, 2016.

6. *Quantifying Retrieval Bias in Web Archive Search*. **Thaer Samar**, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman, and Arjen P. de Vries. International Journal on Digital Libraries, Special Issue on Web Archiving, 2017.

Part II – Integrating Online & Crawled Web- Open Web (live and dynamic) & Crawled Web (archived and static)

1. *Challenges on Combining Open Web and Dataset Evaluation Results: The Case of the Contextual Suggestion Track.* Alejandro Bellogín and **Thaer Samar** and Arjen P. de Vries and Alan Said. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings.
2. *The Strange Case of Reproducibility vs. Representativeness in Contextual Suggestion Test Collections.* **Thaer Samar** and Alejandro Bellogín and Arjen P. de Vries. Information Retrieval Journal, 2015.
3. *Better Contextual Suggestions in ClueWeb12 Using Domain Knowledge Inferred from The Open Web.* **Thaer Samar** and Arjen P. de Vries and Alejandro Bellogín. Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014.
4. *CWI and TU Delft Notebook TREC 2013: Contextual Suggestion, Federated Web Search, KBA, and Web Tracks .* Alejandro Bellogín and Gebrekirstos G. Gebremeskel and Jiyin He andAlan Said and **Thaer Samar** and Arjen P. de Vries and Jimmy Lin and Jeroen B. P. Vuurens. Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013.
5. *Improving Contextual Suggestions using Open Web Domain Knowledge.* **Thaer Samar** and Alejandro Bellogín and Arjen P. de Vries. Proceedings of the International Workshop on Social Personalisation & Search, SPS 2015, co-located with the 38th Annual ACM SIGIR Conference (SIGIR 2015), Santiago de Chile, Chile, August 9-13, 2015.

Other Publications

1. *Sprint methods for web archive research.* Hugo C. Huurdeman and Anat Ben-David and **Thaer Samar**. Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013.
2. *Column Stores as an IR Prototyping Tool.* Hannes Mühleisen and **Thaer Samar** and Jimmy Lin and Arjen P. de Vries. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings.

3. *Old dogs are great at new tricks: column stores for ir prototyping.* Hannes Mühleisen and **Thaer Samar** and Jimmy Lin and Arjen P. de Vries. The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014.

4. *Impact Analysis of OCR Quality on Research Tasks in Digital Archives.* Myriam Traub, **Thaer Samar**, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries and Lynda Hardman. Joint Conference on Digital Libraries JCDL 2016, Newark, New Jersey, USA, June 19-23,2016.

LIST OF FIGURES

Figure 1 **Thesis Structure.** 6

Figure 2 The Internet Archive Wayback Machine URL-based search interface. The screen shot was taken on September 1 2016 at 11:55 AM (CEST). 13

Figure 3 Crawls of `https://www.cwi.nl/` returned by the Internet Archive Wayback Machine. 14

Figure 4 Mementos of `https://www.cwi.nl/` returned by the Memento Time Travel portal. One Memento per archive which holds at least one Memento. 15

Figure 5 'Layers' of contents of the Dutch Web Archive (2012) 31

Figure 6 Number of unique source pages (based on MD5 hash) compared to subset coverage 35

Figure 7 Number of unique words compared to subset coverage 36

Figure 8 Anchor Text frequency distribution of KB_{links} in log scale representation. 67

Figure 9 Distribution of number of versions of documents in the Dutch Web archive collection in log scale representation. 81

Figure 10 Distribution of retrievability scores $r(d)$ for *BM25* based on all documents in the collection at $c = 100$. 86

Figure 11 Distribution of retrievability scores $r(d)$ for *BM25* based on all documents in the collection at $c = 1,000$. 87

Figure 12 **Retrievability scores inequality** among documents in the **entire collection** visualized with Lorenz Curve. 89

Figure 13 **Retrievability scores inequality** among documents in the ***3Models_union_c*** visualized with Lorenz Curve. 90

Figure 14 Number of versions vs. retrievability score, for the *BM25* model. 94

Figure 15 Lorenz curves visualizing the inequality of retrievability scores induced by *BM25* for three scenarios; exact match (green), any match (blue), and cluster match (red). 95

Figure 16 Illustration of the relation between pools and the source of the documents. Subset 1represents the documents in the *Open Web* pool and were found in *ClueWeb12* collection but do not exist in the *ClueWeb12* pool (this subset is used to expand the *ClueWeb12* pool). Subset 2represents the overlap between the *Open Web* pool and *ClueWeb12* pool, documents in this subset were double judged (we use this subset to show the bias between *Open Web* and *ClueWeb12* results). 110

Figure 17 Judgments (document relevance) histogram of documents from *Open Web* (left) and from *ClueWeb12* (right). **CS 2013** 117

Figure 18 Judgments (document relevance) histogram of documents from *Open Web* runs (`left`) and *ClueWeb12* runs (`right`). **CS 2014** 117

Figure 19 Judgments histogram of documents from *Open Web* qrels which exist in *ClueWeb12* collection for CS 2013 (left) and CS 2014 (right) 118

Figure 20 Judgments histogram of documents that exist in both *Open Web* qrels and *ClueWeb12* qrels. Figure on the (left) shows how these documents were judged as *Open Web* URLs, while the figure on the (right) shows how the same documents were judged as *ClueWeb12* documents. **CS 2013** 119

Figure 21 Judgments histogram of documents that exist in both *Open Web* qrels and *ClueWeb12* qrels. Figure on the (left) shows how these documents were judged as *Open Web* URLs, while the figure on the (right) shows how the same documents were judged as *ClueWeb12* documents. **CS 2014** 119

Figure 22 Distribution of the document length in words for the **GeographicFiltered** (left) and *Attractions* (right) sub-collections. Note the different range in the X axis. 135

Figure 23 Screenshots of a document retrieved by the **GeographicFiltered** sub-collection (left) and by the *Attractions* sub-collection (right). The document in the left (clueweb12-0202wb-00-19744) was rated in average with a value of 1.9, whereas the one in the right (clueweb12-0200tw-67-19011) with a 3. 136

LIST OF TABLES

Table 1 Number of documents per year 27
Table 2 Unique archived pages (2012) 30
Table 3 Unique archived hosts, domains & TLDs 30
Table 4 Coverage in archive 30
Table 5 Unarchived *aura* unique pages (2012) 31
Table 6 Unarchived unique hosts, domains & TLDs 32
Table 7 Unarchived *aura* coverage (2012) 32
Table 8 Unarchived *aura* filetypes 32
Table 9 TLD distribution 33
Table 10 Coverage of most popular Dutch sites (*Alexa position*) 34
Table 11 Link types 35
Table 12 Target structure distribution 37
Table 13 Sample aggregated Anchor Text words 38
Table 14 Mean Reciprocal Rank (MRR) 40
Table 15 Success rates (target page in top 10) 41
Table 16 Division based on indegree of unique hosts 42
Table 17 Number of seeds and archived objects over the years 46
Table 18 Percentage of new target hosts over the years considering the top 1,000, 5,000, and 10,000 hosts. 49
Table 19 Top ranked hosts over the years. 50
Table 20 Top hosts per month. **02-06/2009** 51
Table 21 Top hosts per month. **07-12/2009** 52
Table 22 Absolute count and percentage of Anchor Text per year that has a Wikipedia title match at different thresholds. 53
Table 23 List of Anchor Text in the top-1k of 2012, that have no matching Wikipedia title. 54
Table 24 List of Anchor Text the top-1k of 2012 which have matching Wikipedia titles. 54
Table 25 Number of unique links in each dataset. 61
Table 26 Number of unique topics per source. 62

Table 27 **Analysis of hosts:** For both the target and source pages, we present the absolute count of unique hosts (*first row*), the fraction of hosts from KB_{links} that were found in the corresponding dataset in column header (*second row*), and present information about target hosts that has been crawled in each link dataset (absolute count and percentage). 63

Table 28 **TLDs of target pages:** The count of unique TLDs, and the top-10 TLDs. 64

Table 29 **Anchor Text summary:** For each link dataset, we present the number of unique Anchor Text, and the overlap of Anchor Text between KB_{links} and the corresponding dataset. Considering all Anchor Text in KB_{links} (%overlap_all), and by considering Anchor Text used at least twice in KB_{links} (%overlap_GT1). 65

Table 30 **Topic Coverage:** for each link dataset, we present the absolute count and the fraction (%) of found topics in each topic source, where the fraction is the number of matched topics to the total number of topics in the corresponding source. The %*lost* under $CC_{links} \setminus KB_{targets}$ is the relative not found topics, these topics were found in CC_{links} but in $CC_{links} \setminus KB_{targets}$. 66

Table 31 **Unique Topic Coverage in KB_{links}:** in comparison with topics found in other datasets. Under every link dataset x, we present the percentage of topics found in the KB but not found in x (*first column*), and the percentage of topics found in x but not found in KB_{links} (*second column*). 68

Table 32 **The fraction of top-c ranked Anchor Text matching document titles in *WikiStats***, fraction in percentage (*num. matches/c*). In addition to the percentage of matching when all Anchor Text are used (*num. matches/num. all Anchor Text*). 69

Table 33 **The fraction of top-c ranked Anchor Text matching user queries**, same notation as in Table 32. 69

Table 34 Summary of the archived objects over the years, with more details on documents of *text/HTML* content-type. The mean value of number of versions was computed by dividing the total number of document versions crawled per year over the unique number of documents (URLs). The number of original URLs for *All* years is the number of unique URLs in the four years. 80

Table 35 Summary of the query sets. 83

Table 36 Percentage of overlap between the vocabulary of the query sets at different cutoff levels after sorting terms in descending order. 84

Table 37 Query length distribution of queries in the Q_a query set. 84

Table 38 **Gini Coefficients** for all retrieval models with different values of c; all documents in the collection are used for computing the Gini Coefficient. The retrievability score was computed based on the document version granularity. 88

Table 39 **Gini Coefficients** for all retrieval models with different values of c; document versions in the *3Models_union_c* at the corresponding c are considered for computing the Gini Coefficient. The % retrieved is the fraction of retrieved document versions using the models from the whole collection at corresponding c. 91

Table 40 Effectiveness of known-item queries measured by MRR. The first bin consists of the least retrievable documents, while the fourth bin contains the most retrievable documents. An $*$ indicates that the difference between the corresponding bin and the fourth bin is not significant using the Kolmogorov-Smirnov ($p > 0.05$). 92

Table 41 **Gini Coefficients** for all retrieval models based on the two query sets. **Any version**. 93

Table 42 **Gini Coefficients** for all retrieval models based on the two query sets. **Cluster version**. 96

Table 43 **Gini Coefficients** for the three retrieval models based on the two query sets, after embedding the prior based on number of versions with content similarity weight. 98

Table 44 **Retrievability subset analysis** using *BM25* results. For every subset, query set, and c: We present the fraction of retrieved documents from the subset in percentage (num. retrieved / subset size) (**first column**). The mean retrievability score of retrieved documents (**second column**). The fraction of retrieved documents from the corresponding subset to the number of all retrieved (num. retrieved per subset / all retrieved (all subsets)) in percentages (**third column**); sum of the percentage in this column is equal to 100%. 99

Table 45 Query length distribution in the Q_a query set per year. 100

Table 46 Summary of query subsets of Q_a query set. For each subset, we show the number of queries. In parentheses is the number of unique queries in the corresponding subset (year) compared to previous years. For example, the Q_a *_2012* is compared against 2009, 2010, and 2011. For the 2009 subset the percentage of unique Anchor Text is *N/A* as it is the first, and the percentage decreases across the years. 101

Table 47 **Gini Coefficients** for the three models at different c's using different query subsets, using documents in the *3Models_union_c* generated based on running the Q_a query set. 102

Table 48 **Retrievability subset analysis based on time-aware queries** using *BM25* results. The fraction of retrieved documents per year to the total documents retrieved using *BM25* (*%retrieved*). The *%gain* represents the relative percentage of documents that we get per year using the corresponding query set to the % retrieved of the same year using the entire Q_a query set (Table 44). 103

Table 49 Summary of judged documents form the *Open Web* and the *ClueWeb12* collection. The `total` column shows the total number of judged documents, while the `unique` presents the number of unique documents. 114

Table 50 URLs obtained from *Open Web* runs. 115

Table 51 Performance of *Open Web* systems on *Open Web* data vs. their performance on *ClueWeb12* data. Under each metric we present three values: original, replaced, and the relative improvement in effectiveness. The column named `original` presents the performance of submitted runs using the original qrels as provided by the organizers, whereas the column `replaced` shows the performance of modified runs (replacing URLs with their match *ClueWeb12* id and removing URLs with no match) using the expanded qrels. The % of *ClueWeb12* documents in top-5 column presents the percentage of *ClueWeb12* documents in the top-5 after replacing the URLs with their match *ClueWeb12* ids while preserving the ranks. The *ClueWeb12* systems(underlined) are included to show how they perform in comparison with *Open Web* systems evaluated on *ClueWeb12* data. For *ClueWeb12* systems no replacement has been applied, denoted by *n/a* under `replaced` and % of improvement. **CS 2013 systems** 122

Table 52 Performance of *Open Web* systems on *Open Web* data vs. their performance on *ClueWeb12* data. Notation as in Table 51. **CS 2014 systems** 123

Table 53 Number of documents for each part of the **TouristFiltered** subcollection. 131

Table 54 Distribution of *ClueWeb12* documents over URLs depth. 132

Table 55 Distribution of URLs depth over the documents in the *Open Web* qrels, and documents in the *ClueWeb12* qrels. 133

Table 56 Distribution of URLs depth over the documents from *Open Web* qrels that exist in *ClueWeb12* collection. 134

Table 57 Performance of **GeographicFiltered** and **TouristFiltered** runs. Analysis per relevance dimension is considered; description (desc), document (doc), and geographical (geo) relevance. We denote with (all) when desc, doc, and geo relevance are considered. 138

Table 58 Comparison between the two runs based on **GeographicFiltered** and **TouristFiltered** subcollections, by showing the percentage of topics where the **TouristFiltered** subcollection gives better, equal, or worse performance compared to the **GeographicFiltered** subcollection. 138

Table 59 Effect of domain knowledge filters on **TouristFiltered** run performance. Union means adding suggestions from the subset filter shown in column header of current column to the previous one. The percentage shows the relative improvement in effectiveness due to filter. 140

Table 60 Effect of using a prior-probability of relevance on the **GeographicFiltered** run performance. *no prior* means applying the general ranking model with $P(s) = 1$ for documents that pass the geo_filter. 140

Table 61 Language model constructed from relevant and not relevant documents. 141

SIKS DISSERTATION SERIES

1998

1 Johan van den Akker (CWI) *DEGAS: An Active, Temporal Database of Autonomous Objects*

2 Floris Wiesman (UM) *Information Retrieval by Graphically Browsing Meta-Information*

3 Ans Steuten (TUD) *A Contribution to the Linguistic Analysis of Business Conversations*

4 Dennis Breuker (UM) *Memory versus Search in Games*

5 E. W. Oskamp (RUL) *Computerondersteuning bij Straftoemeting*

1999

1 Mark Sloof (VUA) *Physiology of Quality Change Modelling: Automated modelling of*

2 Rob Potharst (EUR) *Classification using decision trees and neural nets*

3 Don Beal (UM) *The Nature of Minimax Search*

4 Jacques Penders (UM) *The practical Art of Moving Physical Objects*

5 Aldo de Moor (KUB) *Empowering Communities: A Method for the Legitimate User-Driven*

6 Niek J. E. Wijngaards (VUA) *Re-design of compositional systems*

7 David Spelt (UT) *Verification support for object database design*

8 Jacques H. J. Lenting (UM) *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism*

2000

1 Frank Niessink (VUA) *Perspectives on Improving Software Maintenance*

2 Koen Holtman (TUe) *Prototyping of CMS Storage Management*

3 Carolien M. T. Metselaar (UvA) *Sociaal-organisatorische gevolgen van kennistechnologie*

4 Geert de Haan (VUA) *ETAG, A Formal Model of Competence Knowledge for User Interface*

5 Ruud van der Pol (UM) *Knowledge-based Query Formulation in Information Retrieval*

6 Rogier van Eijk (UU) *Programming Languages for Agent Communication*

7 Niels Peek (UU) *Decision-theoretic Planning of Clinical Patient Management*

8 Veerle Coupé (EUR) *Sensitivity Analyis of Decision-Theoretic Networks*

9 Florian Waas (CWI) *Principles of Probabilistic Query Optimization*

10 Niels Nes (CWI) *Image Database Management System Design Considerations, Algorithms and Architecture*

11 Jonas Karlsson (CWI) *Scalable Distributed Data Structures for Database Management*

2001

1 Silja Renooij (UU) *Qualitative Approaches to Quantifying Probabilistic Networks*

2 Koen Hindriks (UU) *Agent Programming Languages: Programming with Mental Models*

3 Maarten van Someren (UvA) *Learning as problem solving*

4 Evgueni Smirnov (UM) *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*

5 Jacco van Ossenbruggen (VUA) *Processing Structured Hypermedia: A Matter of Style*

6 Martijn van Welie (VUA) *Task-based User Interface Design*

7 Bastiaan Schonhage (VUA) *Diva: Architectural Perspectives on Information Visualization*

8 Pascal van Eck (VUA) *A Compositional Semantic Structure for Multi-Agent Systems Dynamics*

9 Pieter Jan 't Hoen (RUL) *Towards Distributed Development of Large Object-Oriented Models*

10 Maarten Sierhuis (UvA) *Modeling and Simulating Work Practice*

11 Tom M. van Engers (VUA) *Knowledge Management*

2002

1 Nico Lassing (VUA) *Architecture-Level Modifiability Analysis*

2 Roelof van Zwol (UT) *Modelling and searching web-based document collections*
3 Henk Ernst Blok (UT) *Database Optimization Aspects for Information Retrieval*
4 Juan Roberto Castelo Valdueza (UU) *The Discrete Acyclic Digraph Markov Model in Data Mining*
5 Radu Serban (VUA) *The Private Cyberspace Modeling Electronic*
6 Laurens Mommers (UL) *Applied legal epistemology: Building a knowledge-based ontology of*
7 Peter Boncz (CWI) *Monet: A Next-Generation DBMS Kernel For Query-Intensive*
8 Jaap Gordijn (VUA) *Value Based Requirements Engineering: Exploring Innovative*
9 Willem-Jan van den Heuvel (KUB) *Integrating Modern Business Applications with Objectified Legacy*
10 Brian Sheppard (UM) *Towards Perfect Play of Scrabble*
11 Wouter C. A. Wijngaards (VUA) *Agent Based Modelling of Dynamics: Biological and Organisational Applications*
12 Albrecht Schmidt (UvA) *Processing XML in Database Systems*
13 Hongjing Wu (TUe) *A Reference Architecture for Adaptive Hypermedia Applications*
14 Wieke de Vries (UU) *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*
15 Rik Eshuis (UT) *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*
16 Pieter van Langen (VUA) *The Anatomy of Design: Foundations, Models and Applications*
17 Stefan Manegold (UvA) *Understanding, Modeling, and Improving Main-Memory Database Performance*

2003

1 Heiner Stuckenschmidt (VUA) *Ontology-Based Information Sharing in Weakly Structured Environments*
2 Jan Broersen (VUA) *Modal Action Logics for Reasoning About Reactive Systems*
3 Martijn Schuemie (TUD) *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*
4 Milan Petkovic (UT) *Content-Based Video Retrieval Supported by Database Technology*
5 Jos Lehmann (UvA) *Causation in Artificial Intelligence and Law: A modelling approach*
6 Boris van Schooten (UT) *Development and specification of virtual environments*
7 Machiel Jansen (UvA) *Formal Explorations of Knowledge Intensive Tasks*
8 Yongping Ran (UM) *Repair Based Scheduling*
9 Rens Kortmann (UM) *The resolution of visually guided behaviour*
10 Andreas Lincke (UvT) *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*
11 Simon Keizer (UT) *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*
12 Roeland Ordelman (UT) *Dutch speech recognition in multimedia information retrieval*
13 Jeroen Donkers (UM) *Nosce Hostem: Searching with Opponent Models*
14 Stijn Hoppenbrouwers (KUN) *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*
15 Mathijs de Weerdt (TUD) *Plan Merging in Multi-Agent Systems*
16 Menzo Windhouwer (CWI) *Feature Grammar Systems: Incremental Maintenance of Indexes to Digital Media Warehouses*
17 David Jansen (UT) *Extensions of Statecharts with Probability, Time, and Stochastic Timing*
18 Levente Kocsis (UM) *Learning Search Decisions*

2004

1 Virginia Dignum (UU) *A Model for Organizational Interaction: Based on Agents, Founded in Logic*
2 Lai Xu (UvT) *Monitoring Multi-party Contracts for E-business*
3 Perry Groot (VUA) *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*
4 Chris van Aart (UvA) *Organizational Principles for Multi-Agent Architectures*
5 Viara Popova (EUR) *Knowledge discovery and monotonicity*
6 Bart-Jan Hommes (TUD) *The Evaluation of Business Process Modeling Techniques*

7 Elise Boltjes (UM) *Voorbeeldig onderwijs: voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*
8 Joop Verbeek (UM) *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise*
9 Martin Caminada (VUA) *For the Sake of the Argument: explorations into argument-based reasoning*
10 Suzanne Kabel (UvA) *Knowledge-rich indexing of learning-objects*
11 Michel Klein (VUA) *Change Management for Distributed Ontologies*
12 The Duy Bui (UT) *Creating emotions and facial expressions for embodied agents*
13 Wojciech Jamroga (UT) *Using Multiple Models of Reality: On Agents who Know how to Play*
14 Paul Harrenstein (UU) *Logic in Conflict. Logical Explorations in Strategic Equilibrium*
15 Arno Knobbe (UU) *Multi-Relational Data Mining*
16 Federico Divina (VUA) *Hybrid Genetic Relational Search for Inductive Learning*
17 Mark Winands (UM) *Informed Search in Complex Games*
18 Vania Bessa Machado (UvA) *Supporting the Construction of Qualitative Knowledge Models*
19 Thijs Westerveld (UT) *Using generative probabilistic models for multimedia retrieval*
20 Madelon Evers (Nyenrode) *Learning from Design: facilitating multidisciplinary design teams*

2005

1 Floor Verdenius (UvA) *Methodological Aspects of Designing Induction-Based Applications*
2 Erik van der Werf (UM) *AI techniques for the game of Go*
3 Franc Grootjen (RUN) *A Pragmatic Approach to the Conceptualisation of Language*
4 Nirvana Meratnia (UT) *Towards Database Support for Moving Object data*
5 Gabriel Infante-Lopez (UvA) *Two-Level Probabilistic Grammars for Natural Language Parsing*
6 Pieter Spronck (UM) *Adaptive Game AI*
7 Flavius Frasincar (TUe) *Hypermedia Presentation Generation for Semantic Web Information Systems*
8 Richard Vdovjak (TUe) *A Model-driven Approach for Building Distributed Ontology-based Web Applications*
9 Jeen Broekstra (VUA) *Storage, Querying and Inferencing for Semantic Web Languages*
10 Anders Bouwer (UvA) *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*
11 Elth Ogston (VUA) *Agent Based Matchmaking and Clustering: A Decentralized Approach to Search*
12 Csaba Boer (EUR) *Distributed Simulation in Industry*
13 Fred Hamburg (UL) *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*
14 Borys Omelayenko (VUA) *Web-Service configuration on the Semantic Web: Exploring how semantics meets pragmatics*
15 Tibor Bosse (VUA) *Analysis of the Dynamics of Cognitive Processes*
16 Joris Graaumans (UU) *Usability of XML Query Languages*
17 Boris Shishkov (TUD) *Software Specification Based on Re-usable Business Components*
18 Danielle Sent (UU) *Test-selection strategies for probabilistic networks*
19 Michel van Dartel (UM) *Situated Representation*
20 Cristina Coteanu (UL) *Cyber Consumer Law, State of the Art and Perspectives*
21 Wijnand Derks (UT) *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*

2006

1 Samuil Angelov (TUe) *Foundations of B2B Electronic Contracting*
2 Cristina Chisalita (VUA) *Contextual issues in the design and use of information technology in organizations*
3 Noor Christoph (UvA) *The role of metacognitive skills in learning to solve problems*
4 Marta Sabou (VUA) *Building Web Service Ontologies*
5 Cees Pierik (UU) *Validation Techniques for Object-Oriented Proof Outlines*

6 Ziv Baida (VUA) *Software-aided Service Bundling: Intelligent Methods & Tools for Graphical Service Modeling*

7 Marko Smiljanic (UT) *XML schema matching: balancing efficiency and effectiveness by means of clustering*

8 Eelco Herder (UT) *Forward, Back and Home Again: Analyzing User Behavior on the Web*

9 Mohamed Wahdan (UM) *Automatic Formulation of the Auditor's Opinion*

10 Ronny Siebes (VUA) *Semantic Routing in Peer-to-Peer Systems*

11 Joeri van Ruth (UT) *Flattening Queries over Nested Data Types*

12 Bert Bongers (VUA) *Interactivation: Towards an e-cology of people, our technological environment, and the arts*

13 Henk-Jan Lebbink (UU) *Dialogue and Decision Games for Information Exchanging Agents*

14 Johan Hoorn (VUA) *Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change*

15 Rainer Malik (UU) *CONAN: Text Mining in the Biomedical Domain*

16 Carsten Riggelsen (UU) *Approximation Methods for Efficient Learning of Bayesian Networks*

17 Stacey Nagata (UU) *User Assistance for Multitasking with Interruptions on a Mobile Device*

18 Valentin Zhizhkun (UvA) *Graph transformation for Natural Language Processing*

19 Birna van Riemsdijk (UU) *Cognitive Agent Programming: A Semantic Approach*

20 Marina Velikova (UvT) *Monotone models for prediction in data mining*

21 Bas van Gils (RUN) *Aptness on the Web*

22 Paul de Vrieze (RUN) *Fundaments of Adaptive Personalisation*

23 Ion Juvina (UU) *Development of Cognitive Model for Navigating on the Web*

24 Laura Hollink (VUA) *Semantic Annotation for Retrieval of Visual Resources*

25 Madalina Drugan (UU) *Conditional log-likelihood MDL and Evolutionary MCMC*

26 Vojkan Mihajlovic (UT) *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*

27 Stefano Bocconi (CWI) *Vox Populi: generating video documentaries from semantically annotated media repositories*

28 Borkur Sigurbjornsson (UvA) *Focused Information Access using XML Element Retrieval*

2007

1 Kees Leune (UvT) *Access Control and Service-Oriented Architectures*

2 Wouter Teepe (RUG) *Reconciling Information Exchange and Confidentiality: A Formal Approach*

3 Peter Mika (VUA) *Social Networks and the Semantic Web*

4 Jurriaan van Diggelen (UU) *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*

5 Bart Schermer (UL) *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*

6 Gilad Mishne (UvA) *Applied Text Analytics for Blogs*

7 Natasa Jovanovic' (UT) *To Whom It May Concern: Addressee Identification in Face-to-Face Meetings*

8 Mark Hoogendoorn (VUA) *Modeling of Change in Multi-Agent Organizations*

9 David Mobach (VUA) *Agent-Based Mediated Service Negotiation*

10 Huib Aldewereld (UU) *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*

11 Natalia Stash (TUe) *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*

12 Marcel van Gerven (RUN) *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*

13 Rutger Rienks (UT) *Meetings in Smart Environments: Implications of Progressing Technology*

14 Niek Bergboer (UM) *Context-Based Image Analysis*

15 Joyca Lacroix (UM) *NIM: a Situated Computational Memory Model*

16 Davide Grossi (UU) *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*

17 Theodore Charitos (UU) *Reasoning with Dynamic Networks in Practice*

18 Bart Orriens (UvT) *On the development an management of adaptive business collaborations*

19 David Levy (UM) *Intimate relationships with artificial partners*

20 Slinger Jansen (UU) *Customer Configuration Updating in a Software Supply Network*

21 Karianne Vermaas (UU) *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*

22 Zlatko Zlatev (UT) *Goal-oriented design of value and process models from patterns*

23 Peter Barna (TUe) *Specification of Application Logic in Web Information Systems*

24 Georgina Ramírez Camps (CWI) *Structural Features in XML Retrieval*

25 Joost Schalken (VUA) *Empirical Investigations in Software Process Improvement*

2008

1 Katalin Boer-Sorbán (EUR) *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*

2 Alexei Sharpanskykh (VUA) *On Computer-Aided Methods for Modeling and Analysis of Organizations*

3 Vera Hollink (UvA) *Optimizing hierarchical menus: a usage-based approach*

4 Ander de Keijzer (UT) *Management of Uncertain Data: towards unattended integration*

5 Bela Mutschler (UT) *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*

6 Arjen Hommersom (RUN) *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*

7 Peter van Rosmalen (OU) *Supporting the tutor in the design and support of adaptive e-learning*

8 Janneke Bolt (UU) *Bayesian Networks: Aspects of Approximate Inference*

9 Christof van Nimwegen (UU) *The paradox of the guided user: assistance can be counter-effective*

10 Wauter Bosma (UT) *Discourse oriented summarization*

11 Vera Kartseva (VUA) *Designing Controls for Network Organizations: A Value-Based Approach*

12 Jozsef Farkas (RUN) *A Semiotically Oriented Cognitive Model of Knowledge Representation*

13 Caterina Carraciolo (UvA) *Topic Driven Access to Scientific Handbooks*

14 Arthur van Bunningen (UT) *Context-Aware Querying: Better Answers with Less Effort*

15 Martijn van Otterlo (UT) *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains*

16 Henriette van Vugt (VUA) *Embodied agents from a user's perspective*

17 Martin Op 't Land (TUD) *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*

18 Guido de Croon (UM) *Adaptive Active Vision*

19 Henning Rode (UT) *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*

20 Rex Arendsen (UvA) *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven*

21 Krisztian Balog (UvA) *People Search in the Enterprise*

22 Henk Koning (UU) *Communication of IT-Architecture*

23 Stefan Visscher (UU) *Bayesian network models for the management of ventilator-associated pneumonia*

24 Zharko Aleksovski (VUA) *Using background knowledge in ontology matching*

25 Geert Jonker (UU) *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*

26 Marijn Huijbregts (UT) *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*

27 Hubert Vogten (OU) *Design and Implementation Strategies for IMS Learning Design*

28 Ildiko Flesch (RUN) *On the Use of Independence Relations in Bayesian Networks*

29 Dennis Reidsma (UT) *Annotations and Subjective Machines: Of Annotators, Embodied Agents, Users, and Other Humans*

30 Wouter van Atteveldt (VUA) *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*

31 Loes Braun (UM) *Pro-Active Medical Information Retrieval*

32 Trung H. Bui (UT) *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes*

33 Frank Terpstra (UvA) *Scientific Workflow Design: theoretical and practical issues*

34 Jeroen de Knijf (UU) *Studies in Frequent Tree Mining*

35 Ben Torben Nielsen (UvT) *Dendritic morphologies: function shapes structure*

2009

1 Rasa Jurgelenaite (RUN) *Symmetric Causal Independence Models*

2 Willem Robert van Hage (VUA) *Evaluating Ontology-Alignment Techniques*

3 Hans Stol (UvT) *A Framework for Evidence-based Policy Making Using IT*

4 Josephine Nabukenya (RUN) *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

5 Sietse Overbeek (RUN) *Bridging Supply and Demand for Knowledge Intensive Tasks: Based on Knowledge, Cognition, and Quality*

6 Muhammad Subianto (UU) *Understanding Classification*

7 Ronald Poppe (UT) *Discriminative Vision-Based Recovery and Recognition of Human Motion*

8 Volker Nannen (VUA) *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

9 Benjamin Kanagwa (RUN) *Design, Discovery and Construction of Service-oriented Systems*

10 Jan Wielemaker (UvA) *Logic programming for knowledge-intensive interactive applications*

11 Alexander Boer (UvA) *Legal Theory, Sources of Law & the Semantic Web*

12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin) *Operating Guidelines for Services*

13 Steven de Jong (UM) *Fairness in Multi-Agent Systems*

14 Maksym Korotkiy (VUA) *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*

15 Rinke Hoekstra (UvA) *Ontology Representation: Design Patterns and Ontologies that Make Sense*

16 Fritz Reul (UvT) *New Architectures in Computer Chess*

17 Laurens van der Maaten (UvT) *Feature Extraction from Visual Data*

18 Fabian Groffen (CWI) *Armada, An Evolving Database System*

19 Valentin Robu (CWI) *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

20 Bob van der Vecht (UU) *Adjustable Autonomy: Controling Influences on Decision Making*

21 Stijn Vanderlooy (UM) *Ranking and Reliable Classification*

22 Pavel Serdyukov (UT) *Search For Expertise: Going beyond direct evidence*

23 Peter Hofgesang (VUA) *Modelling Web Usage in a Changing Environment*

24 Annerieke Heuvelink (VUA) *Cognitive Models for Training Simulations*

25 Alex van Ballegooij (CWI) *RAM: Array Database Management through Relational Mapping*

26 Fernando Koch (UU) *An Agent-Based Model for the Development of Intelligent Mobile Services*

27 Christian Glahn (OU) *Contextual Support of social Engagement and Reflection on the Web*

28 Sander Evers (UT) *Sensor Data Management with Probabilistic Models*

29 Stanislav Pokraev (UT) *Model-Driven Semantic Integration of Service-Oriented Applications*

30 Marcin Zukowski (CWI) *Balancing vectorized query execution with bandwidth-optimized storage*

31 Sofiya Katrenko (UvA) *A Closer Look at Learning Relations from Text*

32 Rik Farenhorst (VUA) *Architectural Knowledge Management: Supporting Architects and Auditors*

33 Khiet Truong (UT) *How Does Real Affect Affect Affect Recognition In Speech?*

34 Inge van de Weerd (UU) *Advancing in Software Product Management: An Incremental Method Engineering Approach*

35 Wouter Koelewijn (UL) *Privacy en Politiegegevens: Over geautomatiseerde normatieve informatie-uitwisseling*

36 Marco Kalz (OUN) *Placement Support for Learners in Learning Networks*

37 Hendrik Drachsler (OUN) *Navigation Support for Learners in Informal Learning Networks*

38 Riina Vuorikari (OU) *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*

39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin) *Service Substitution: A Behavioral Approach Based on Petri Nets*

40 Stephan Raaijmakers (UvT) *Multinomial Language Learning: Investigations into the Geometry of Language*

41 Igor Berezhnyy (UvT) *Digital Analysis of Paintings*

42 Toine Bogers (UvT) *Recommender Systems for Social Bookmarking*

43 Virginia Nunes Leal Franqueira (UT) *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*

44 Roberto Santana Tapia (UT) *Assessing Business-IT Alignment in Networked Organizations*

45 Jilles Vreeken (UU) *Making Pattern Mining Useful*

46 Loredana Afanasiev (UvA) *Querying XML: Benchmarks and Recursion*

2010

1 Matthijs van Leeuwen (UU) *Patterns that Matter*

2 Ingo Wassink (UT) *Work flows in Life Science*

3 Joost Geurts (CWI) *A Document Engineering Model and Processing Framework for Multimedia documents*

4 Olga Kulyk (UT) *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*

5 Claudia Hauff (UT) *Predicting the Effectiveness of Queries and Retrieval Systems*

6 Sander Bakkes (UvT) *Rapid Adaptation of Video Game AI*

7 Wim Fikkert (UT) *Gesture interaction at a Distance*

8 Krzysztof Siewicz (UL) *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*

9 Hugo Kielman (UL) *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*

10 Rebecca Ong (UL) *Mobile Communication and Protection of Children*

11 Adriaan Ter Mors (TUD) *The world according to MARP: Multi-Agent Route Planning*

12 Susan van den Braak (UU) *Sensemaking software for crime analysis*

13 Gianluigi Folino (RUN) *High Performance Data Mining using Bio-inspired techniques*

14 Sander van Splunter (VUA) *Automated Web Service Reconfiguration*

15 Lianne Bodenstaff (UT) *Managing Dependency Relations in Inter-Organizational Models*

16 Sicco Verwer (TUD) *Efficient Identification of Timed Automata, theory and practice*

17 Spyros Kotoulas (VUA) *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*

18 Charlotte Gerritsen (VUA) *Caught in the Act: Investigating Crime by Agent-Based Simulation*

19 Henriette Cramer (UvA) *People's Responses to Autonomous and Adaptive Systems*

20 Ivo Swartjes (UT) *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*

21 Harold van Heerde (UT) *Privacy-aware data management by means of data degradation*

22 Michiel Hildebrand (CWI) *End-user Support for Access to Heterogeneous Linked Data*

23 Bas Steunebrink (UU) *The Logical Structure of Emotions*

24 Zulfiqar Ali Memon (VUA) *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*

25 Ying Zhang (CWI) *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*

26 Marten Voulon (UL) *Automatisch contracteren*

27 Arne Koopman (UU) *Characteristic Relational Patterns*

28 Stratos Idreos (CWI) *Database Cracking: Towards Auto-tuning Database Kernels*

29 Marieke van Erp (UvT) *Accessing Natural History: Discoveries in data cleaning, structuring, and retrieval*

30 Victor de Boer (UvA) *Ontology Enrichment from Heterogeneous Sources on the Web*

31 Marcel Hiel (UvT) *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*

32 Robin Aly (UT) *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*

33 Teduh Dirgahayu (UT) *Interaction Design in Service Compositions*

34 Dolf Trieschnigg (UT) *Proof of Concept: Concept-based Biomedical Information Retrieval*

35 Jose Janssen (OU) *Paving the Way for Lifelong Learning: Facilitating competence development through a learning path specification*

36 Niels Lohmann (TUe) *Correctness of services and their composition*

37 Dirk Fahland (TUe) *From Scenarios to components*

38 Ghazanfar Farooq Siddiqui (VUA) *Integrative modeling of emotions in virtual agents*

39 Mark van Assem (VUA) *Converting and Integrating Vocabularies for the Semantic Web*

40 Guillaume Chaslot (UM) *Monte-Carlo Tree Search*

41 Sybren de Kinderen (VUA) *Needs-driven service bundling in a multi-supplier setting: the computational e3-service approach*

42 Peter van Kranenburg (UU) *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*

43 Pieter Bellekens (TUe) *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*

44 Vasilios Andrikopoulos (UvT) *A theory and model for the evolution of software services*

45 Vincent Pijpers (VUA) *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

46 Chen Li (UT) *Mining Process Model Variants: Challenges, Techniques, Examples*

47 Jahn-Takeshi Saito (UM) *Solving difficult game positions*

48 Bouke Huurnink (UvA) *Search in Audiovisual Broadcast Archives*

49 Alia Khairia Amin (CWI) *Understanding and supporting information seeking tasks in multiple sources*

50 Peter-Paul van Maanen (VUA) *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*

51 Edgar Meij (UvA) *Combining Concepts and Language Models for Information Access*

2011

1 Botond Cseke (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

2 Nick Tinnemeier (UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

3 Jan Martijn van der Werf (TUe) *Compositional Design and Verification of Component-Based Information Systems*

4 Hado van Hasselt (UU) *Insights in Reinforcement Learning: Formal analysis and empirical evaluation of temporal-difference*

5 Base van der Raadt (VUA) *Enterprise Architecture Coming of Age: Increasing the Performance of an Emerging Discipline*

6 Yiwen Wang (TUe) *Semantically-Enhanced Recommendations in Cultural Heritage*

7 Yujia Cao (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*

8 Nieske Vergunst (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*

9 Tim de Jong (OU) *Contextualised Mobile Media for Learning*

10 Bart Bogaert (UvT) *Cloud Content Contention*

11 Dhaval Vyas (UT) *Designing for Awareness: An Experience-focused HCI Perspective*

12 Carmen Bratosin (TUe) *Grid Architecture for Distributed Process Mining*

13 Xiaoyu Mao (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

14 Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*

15 Marijn Koolen (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

16 Maarten Schadd (UM) *Selective Search in Games of Different Complexity*

17 Jiyin He (UvA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*
18 Mark Ponsen (UM) *Strategic Decision-Making in complex games*
19 Ellen Rusman (OU) *The Mind ' s Eye on Personal Profiles*
20 Qing Gu (VUA) *Guiding service-oriented software engineering: A view-based approach*
21 Linda Terlouw (TUD) *Modularization and Specification of Service-Oriented Systems*
22 Junte Zhang (UvA) *System Evaluation of Archival Description and Access*
23 Wouter Weerkamp (UvA) *Finding People and their Utterances in Social Media*
24 Herwin van Welbergen (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*
25 Syed Waqar ul Qounain Jaffry (VUA) *Analysis and Validation of Models for Trust Dynamics*
26 Matthijs Aart Pontier (VUA) *Virtual Agents for Human Communication: Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*
27 Aniel Bhulai (VUA) *Dynamic website optimization through autonomous management of design patterns*
28 Rianne Kaptein (UvA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*
29 Faisal Kamiran (TUe) *Discrimination-aware Classification*
30 Egon van den Broek (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*
31 Ludo Waltman (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*
32 Nees-Jan van Eck (EUR) *Methodological Advances in Bibliometric Mapping of Science*
33 Tom van der Weide (UU) *Arguing to Motivate Decisions*
34 Paolo Turrini (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*
35 Maaike Harbers (UU) *Explaining Agent Behavior in Virtual Training*
36 Erik van der Spek (UU) *Experiments in serious game design: a cognitive approach*
37 Adriana Burlutiu (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*
38 Nyree Lemmens (UM) *Bee-inspired Distributed Optimization*
39 Joost Westra (UU) *Organizing Adaptation using Agents in Serious Games*
40 Viktor Clerc (VUA) *Architectural Knowledge Management in Global Software Development*
41 Luan Ibraimi (UT) *Cryptographically Enforced Distributed Data Access Control*
42 Michal Sindlar (UU) *Explaining Behavior through Mental State Attribution*
43 Henk van der Schuur (UU) *Process Improvement through Software Operation Knowledge*
44 Boris Reuderink (UT) *Robust Brain-Computer Interfaces*
45 Herman Stehouwer (UvT) *Statistical Language Models for Alternative Sequence Selection*
46 Beibei Hu (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*
47 Azizi Bin Ab Aziz (VUA) *Exploring Computational Models for Intelligent Support of Persons with Depression*
48 Mark Ter Maat (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*
49 Andreea Niculescu (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

1 Terry Kakeeto (UvT) *Relationship Marketing for SMEs in Uganda*
2 Muhammad Umair (VUA) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*
3 Adam Vanya (VUA) *Supporting Architecture Evolution by Mining Software Repositories*
4 Jurriaan Souer (UU) *Development of Content Management System-based Web Applications*
5 Marijn Plomp (UU) *Maturing Interorganisational Information Systems*
6 Wolfgang Reinhardt (OU) *Awareness Support for Knowledge Workers in Research Networks*

7 Rianne van Lambalgen (VUA) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

8 Gerben de Vries (UvA) *Kernel Methods for Vessel Trajectories*

9 Ricardo Neisse (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*

10 David Smits (TUe) *Towards a Generic Distributed Adaptive Hypermedia Environment*

11 J. C. B. Rantham Prabhakara (TUe) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

12 Kees van der Sluijs (TUe) *Model Driven Design and Data Integration in Semantic Web Information Systems*

13 Suleman Shahid (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

14 Evgeny Knutov (TUe) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

15 Natalie van der Wal (VUA) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*

16 Fiemke Both (VUA) *Helping people by understanding them: Ambient Agents supporting task execution and depression treatment*

17 Amal Elgammal (UvT) *Towards a Comprehensive Framework for Business Process Compliance*

18 Eltjo Poort (VUA) *Improving Solution Architecting Practices*

19 Helen Schonenberg (TUe) *What's Next? Operational Support for Business Process Execution*

20 Ali Bahramisharif (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

21 Roberto Cornacchia (TUD) *Querying Sparse Matrices for Information Retrieval*

22 Thijs Vis (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

23 Christian Muehl (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

24 Laurens van der Werff (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

25 Silja Eckartz (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

26 Emile de Maat (UvA) *Making Sense of Legal Text*

27 Hayrettin Gurkok (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

28 Nancy Pascall (UvT) *Engendering Technology Empowering Women*

29 Almer Tigelaar (UT) *Peer-to-Peer Information Retrieval*

30 Alina Pommeranz (TUD) *Designing Human-Centered Systems for Reflective Decision Making*

31 Emily Bagarukayo (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

32 Wietske Visser (TUD) *Qualitative multi-criteria preference representation and reasoning*

33 Rory Sie (OUN) *Coalitions in Cooperation Networks (COCOON)*

34 Pavol Jancura (RUN) *Evolutionary analysis in PPI networks and applications*

35 Evert Haasdijk (VUA) *Never Too Old To Learn: On-line Evolution of Controllers in Swarm- and Modular Robotics*

36 Denis Ssebugwawo (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*

37 Agnes Nakakawa (RUN) *A Collaboration Process for Enterprise Architecture Creation*

38 Selmar Smit (VUA) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

39 Hassan Fatemi (UT) *Risk-aware design of value and coordination networks*

40 Agus Gunawan (UvT) *Information Access for SMEs in Indonesia*

41 Sebastian Kelle (OU) *Game Design Patterns for Learning*

42 Dominique Verpoorten (OU) *Reflection Amplifiers in self-regulated Learning*

43 Anna Tordai (VUA) *On Combining Alignment Techniques*

44 Benedikt Kratz (UvT) *A Model and Language for Business-aware Transactions*

45 Simon Carter (UvA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

46 Manos Tsagkias (UvA) *Mining Social Media: Tracking Content and Predicting Behavior*

47 Jorn Bakker (TUe) *Handling Abrupt Changes in Evolving Time-series Data*

48 Michael Kaisers (UM) *Learning against Learning: Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

49 Steven van Kervel (TUD) *Ontologogy driven Enterprise Information Systems Engineering*

50 Jeroen de Jong (TUD) *Heuristics in Dynamic Sceduling: a practical framework with a case study in elevator dispatching*

2013

1 Viorel Milea (EUR) *News Analytics for Financial Decision Support*

2 Erietta Liarou (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*

3 Szymon Klarman (VUA) *Reasoning with Contexts in Description Logics*

4 Chetan Yadati (TUD) *Coordinating autonomous planning and scheduling*

5 Dulce Pumareja (UT) *Groupware Requirements Evolutions Patterns*

6 Romulo Goncalves (CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

7 Giel van Lankveld (UvT) *Quantifying Individual Player Differences*

8 Robbert-Jan Merk (VUA) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

9 Fabio Gori (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*

10 Jeewanie Jayasinghe Arachchige (UvT) *A Unified Modeling Framework for Service Design*

11 Evangelos Pournaras (TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*

12 Marian Razavian (VUA) *Knowledge-driven Migration to Services*

13 Mohammad Safiri (UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

14 Jafar Tanha (UvA) *Ensemble Approaches to Semi-Supervised Learning Learning*

15 Daniel Hennes (UM) *Multiagent Learning: Dynamic Games and Applications*

16 Eric Kok (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*

17 Koen Kok (VUA) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

18 Jeroen Janssens (UvT) *Outlier Selection and One-Class Classification*

19 Renze Steenhuizen (TUD) *Coordinated Multi-Agent Planning and Scheduling*

20 Katja Hofmann (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*

21 Sander Wubben (UvT) *Text-to-text generation by monolingual machine translation*

22 Tom Claassen (RUN) *Causal Discovery and Logic*

23 Patricio de Alencar Silva (UvT) *Value Activity Monitoring*

24 Haitham Bou Ammar (UM) *Automated Transfer in Reinforcement Learning*

25 Agnieszka Anna Latoszek-Berendsen (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

26 Alireza Zarghami (UT) *Architectural Support for Dynamic Homecare Service Provisioning*

27 Mohammad Huq (UT) *Inference-based Framework Managing Data Provenance*

28 Frans van der Sluis (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

29 Iwan de Kok (UT) *Listening Heads*

30 Joyce Nakatumba (TUe) *Resource-Aware Business Process Management: Analysis and Support*

31 Dinh Khoa Nguyen (UvT) *Blueprint Model and Language for Engineering Cloud Applications*

32 Kamakshi Rajagopal (OUN) *Networking For Learning: The role of Networking in a Lifelong Learner's Professional Development*

33 Qi Gao (TUD) *User Modeling and Personalization in the Microblogging Sphere*

34 Kien Tjin-Kam-Jet (UT) *Distributed Deep Web Search*

35 Abdallah El Ali (UvA) *Minimal Mobile Human Computer Interaction*

36 Than Lam Hoang (TUe) *Pattern Mining in Data Streams*

37 Dirk Börner (OUN) *Ambient Learning Displays*

38 Eelco den Heijer (VUA) *Autonomous Evolutionary Art*
39 Joop de Jong (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*
40 Pim Nijssen (UM) *Monte-Carlo Tree Search for Multi-Player Games*
41 Jochem Liem (UvA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*
42 Léon Planken (TUD) *Algorithms for Simple Temporal Reasoning*
43 Marc Bron (UvA) *Exploration and Contextualization through Interaction and Concepts*

2014

1 Nicola Barile (UU) *Studies in Learning Monotone Models from Data*
2 Fiona Tuliyano (RUN) *Combining System Dynamics with a Domain Modeling Method*
3 Sergio Raul Duarte Torres (UT) *Information Retrieval for Children: Search Behavior and Solutions*
4 Hanna Jochmann-Mannak (UT) *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*
5 Jurriaan van Reijsen (UU) *Knowledge Perspectives on Advancing Dynamic Capability*
6 Damian Tamburri (VUA) *Supporting Networked Software Development*
7 Arya Adriansyah (TUe) *Aligning Observed and Modeled Behavior*
8 Samur Araujo (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*
9 Philip Jackson (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*
10 Ivan Salvador Razo Zapata (VUA) *Service Value Networks*
11 Janneke van der Zwaan (TUD) *An Empathic Virtual Buddy for Social Support*
12 Willem van Willigen (VUA) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*
13 Arlette van Wissen (VUA) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*
14 Yangyang Shi (TUD) *Language Models With Meta-information*
15 Natalya Mogles (VUA) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*
16 Krystyna Milian (VUA) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*
17 Kathrin Dentler (VUA) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*
18 Mattijs Ghijsen (UvA) *Methods and Models for the Design and Study of Dynamic Agent Organizations*
19 Vinicius Ramos (TUe) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*
20 Mena Habib (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*
21 Kassidy Clark (TUD) *Negotiation and Monitoring in Open Environments*
22 Marieke Peeters (UU) *Personalized Educational Games: Developing agent-supported scenario-based training*
23 Eleftherios Sidirourgos (UvA/CWI) *Space Efficient Indexes for the Big Data Era*
24 Davide Ceolin (VUA) *Trusting Semi-structured Web Data*
25 Martijn Lappenschaar (RUN) *New network models for the analysis of disease interaction*
26 Tim Baarslag (TUD) *What to Bid and When to Stop*
27 Rui Jorge Almeida (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*
28 Anna Chmielowiec (VUA) *Decentralized k-Clique Matching*
29 Jaap Kabbedijk (UU) *Variability in Multi-Tenant Enterprise Software*
30 Peter de Cock (UvT) *Anticipating Criminal Behaviour*
31 Leo van Moergestel (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*
32 Naser Ayat (UvA) *On Entity Resolution in Probabilistic Data*
33 Tesfa Tegegne (RUN) *Service Discovery in eHealth*
34 Christina Manteli (VUA) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*

35 Joost van Ooijen (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

36 Joos Buijs (TUe) *Flexible Evolutionary Algorithms for Mining Structured Process Models*

37 Maral Dadvar (UT) *Experts and Machines United Against Cyberbullying*

38 Danny Plass-Oude Bos (UT) *Making brain-computer interfaces better: improving usability through post-processing*

39 Jasmina Maric (UvT) *Web Communities, Immigration, and Social Capital*

40 Walter Omona (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*

41 Frederic Hogenboom (EUR) *Automated Detection of Financial Events in News Text*

42 Carsten Eijckhof (CWI/TUD) *Contextual Multidimensional Relevance Models*

43 Kevin Vlaanderen (UU) *Supporting Process Improvement using Method Increments*

44 Paulien Meesters (UvT) *Intelligent Blauw: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*

45 Birgit Schmitz (OUN) *Mobile Games for Learning: A Pattern-Based Approach*

46 Ke Tao (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*

47 Shangsong Liang (UvA) *Fusion and Diversification in Information Retrieval*

2015

1 Niels Netten (UvA) *Machine Learning for Relevance of Information in Crisis Response*

2 Faiza Bukhsh (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*

3 Twan van Laarhoven (RUN) *Machine learning for network data*

4 Howard Spoelstra (OUN) *Collaborations in Open Learning Environments*

5 Christoph Bösch (UT) *Cryptographically Enforced Search Pattern Hiding*

6 Farideh Heidari (TUD) *Business Process Quality Computation: Computing Non-Functional Requirements to Improve Business Processes*

7 Maria-Hendrike Peetz (UvA) *Time-Aware Online Reputation Analysis*

8 Jie Jiang (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*

9 Randy Klaassen (UT) *HCI Perspectives on Behavior Change Support Systems*

10 Henry Hermans (OUN) *OpenU: design of an integrated system to support lifelong learning*

11 Yongming Luo (TUe) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*

12 Julie M. Birkholz (VUA) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*

13 Giuseppe Procaccianti (VUA) *Energy-Efficient Software*

14 Bart van Straalen (UT) *A cognitive approach to modeling bad news conversations*

15 Klaas Andries de Graaf (VUA) *Ontology-based Software Architecture Documentation*

16 Changyun Wei (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*

17 André van Cleeff (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*

18 Holger Pirk (CWI) *Waste Not, Want Not!: Managing Relational Data in Asymmetric Memories*

19 Bernardo Tabuenca (OUN) *Ubiquitous Technology for Lifelong Learners*

20 Loïs Vanhée (UU) *Using Culture and Values to Support Flexible Coordination*

21 Sibren Fetter (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*

22 Zhemin Zhu (UT) *Co-occurrence Rate Networks*

23 Luit Gazendam (VUA) *Cataloguer Support in Cultural Heritage*

24 Richard Berendsen (UvA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*

25 Steven Woudenberg (UU) *Bayesian Tools for Early Disease Detection*

26 Alexander Hogenboom (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*

27 Sándor Héman (CWI) *Updating compressed column-stores*

28 Janet Bagorogoza (TiU) *Knowledge Management and High Performance: The Uganda Financial Institutions Model for HPO*

29 Hendrik Baier (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*

30 Kiavash Bahreini (OUN) *Real-time Multimodal Emotion Recognition in E-Learning*

31 Yakup Koç (TUD) *On Robustness of Power Grids*

32 Jerome Gard (UL) *Corporate Venture Management in SMEs*

33 Frederik Schadd (UM) *Ontology Mapping with Auxiliary Resources*

34 Victor de Graaff (UT) *Geosocial Recommender Systems*

35 Junchao Xu (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

1 Syed Saiden Abbas (RUN) *Recognition of Shapes by Humans and Machines*

2 Michiel Christiaan Meulendijk (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*

3 Maya Sappelli (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*

4 Laurens Rietveld (VUA) *Publishing and Consuming Linked Data*

5 Evgeny Sherkhonov (UvA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*

6 Michel Wilson (TUD) *Robust scheduling in an uncertain environment*

7 Jeroen de Man (VUA) *Measuring and modeling negative emotions for virtual training*

8 Matje van de Camp (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*

9 Archana Nottamkandath (VUA) *Trusting Crowdsourced Information on Cultural Artefacts*

10 George Karafotias (VUA) *Parameter Control for Evolutionary Algorithms*

11 Anne Schuth (UvA) *Search Engines that Learn from Their Users*

12 Max Knobbout (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*

13 Nana Baah Gyan (VU) *The Web, Speech Technologies and Rural Development in West Africa: An ICT4D Approach*

14 Ravi Khadka (UU) *Revisiting Legacy Software System Modernization*

15 Steffen Michels (RUN) *Hybrid Probabilistic Logics: Theoretical Aspects, Algorithms and Experiments*

16 Guangliang Li (UvA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*

17 Berend Weel (VUA) *Towards Embodied Evolution of Robot Organisms*

18 Albert Meroño Peñuela (VUA) *Refining Statistical Data on the Web*

19 Julia Efremova (TUe) *Mining Social Structures from Genealogical Data*

20 Daan Odijk (UvA) *Context & Semantics in News & Web Search*

21 Alejandro Moreno Célleri (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*

22 Grace Lewis (VU) *Software Architecture Strategies for Cyber-Foraging Systems*

23 Fei Cai (UvA) *Query Auto Completion in Information Retrieval*

24 Brend Wanders (UT) *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*

25 Julia Kiseleva (TUe) *Using Contextual Information to Understand Searching and Browsing Behavior*

26 Dilhan Thilakarathne (VU) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*

27 Wen Li (TUD) *Understanding Geo-spatial Information on Social Media*

28 Mingxin Zhang (TUD) *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*

29 Nicolas Höning (CWI/TUD) *Peak reduction in decentralised electricity systems - Markets and prices for flexible planning*

30 Ruud Mattheij (UvT) *The Eyes Have It*

31 Mohammadreza Khelghati (UT) *Deep web content monitoring*

32 Eelco Vriezekolk (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*

33 Peter Bloem (UvA) *Single Sample Statistics, exercises in learning from just one example*
34 Dennis Schunselaar (TUe) *Configurable Process Trees: Elicitation, Analysis, and Enactment*
35 Zhaochun Ren (UvA) *Monitoring Social Media: Summarization, Classification and Recommendation*
36 Daphne Karreman (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*
37 Giovanni Sileno (UvA) *Aligning Law and Action: a conceptual and computational inquiry*
38 Andrea Minuto (UT) *MATERIALS THAT MATTER: Smart Materials meet Art & Interaction Design*
39 Merijn Bruijnes (UT) *Believable Suspect Agents: Response and Interpersonal Style Selection for an Artificial Suspect*
40 Christian Detweiler (TUD) *Accounting for Values in Design*
41 Thomas King (TUD) *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*
42 Spyros Martzoukos (UvA) *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*
43 Saskia Koldijk (RUN) *Context-Aware Support for Stress Self-Management: From Theory to Practice*
44 Thibault Sellam (UvA) *Automatic Assistants for Database Exploration*
45 Bram van de Laar (UT) *Experiencing Brain-Computer Interface Control*
46 Jorge Gallego Perez (UT) *Robots to Make you Happy*
47 Christina Weber (UL) *Real-time foresight: Preparedness for dynamic innovation networks*
48 Tanja Buttler (TUD) *Collecting Lessons Learned*
49 Gleb Polevoy (TUD) *Participation and Interaction in Projects. A Game-Theoretic Analysis*
50 Yan Wang (UVT) *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*

2017

1 Jan-Jaap Oerlemans (UL) *Investigating Cybercrime*
2 Sjoerd Timmer (UU) *Designing and Understanding Forensic Bayesian Networks using Argumentation*
3 Daniël Harold Telgen (UU) *Grid Manufacturing: A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*
4 Mrunal Gawade (CWI) *Multi-core Parallelism in a Column-store*
5 Mahdieh Shadi (UvA) *Collaboration Behavior*
6 Damir Vandic (EUR) *Intelligent Information Systems for Web Product Search*
7 Roel Bertens (UU) *Insight in Information: from Abstract to Anomaly*
8 Rob Konijn (VUA) *Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*
9 Dong Nguyen (UT) *Text as Social and Cultural Data: A Computational Perspective on Variation in Text*
10 Robby van Delden (UT) *(Steering) Interactive Play Behavior*
11 Florian Kunneman (RUN) *Modelling patterns of time and emotion in Twitter #anticipointment*
12 Sander Leemans (UT) *Robust Process Mining with Guarantees*
13 Gijs Huisman (UT) *Social Touch Technology: Extending the reach of social touch through haptic technology*
14 Shoshannah Tekofsky (UvT) *You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior*
15 Peter Berck (RUN) *Memory-Based Text Correction*
16 Aleksandr Chuklin (UvA) *Understanding and Modeling Users of Modern Search Engines*
17 Daniel Dimov (UL) *Crowdsourced Online Dispute Resolution*
18 Ridho Reinanda (UVA) *Entity Associations for Search*
19 Jeroen Vuurens (TUD) *Proximity of Terms, Texts and Semantic Vectors in Information Retrieval*
20 Mohammadbashir Sedighi (TUD) *Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility*
21 Jeroen Linssen (UT) *Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)*
22 Sara Magliacane (VU) *Logics for causal inference under uncertainty*
23 David Graus (UVA) *Entities of Interest — Discovery in Digital Traces*
24 Chang Wang (TUD) *Use of Affordances for Efficient Robot Learning*

25 Veruska Zamborlini (VUA) *Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search*

26 Merel Jung (UT) *Socially intelligent robots that understand and respond to human touch*

27 Michiel Joosse (UT) *Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors*

28 John Klein (VU) *Architecture Practices for Complex Contexts*

29 Adel Alhuraibi (UVT) *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT*

30 Wilma Latuny (UVT) *The Power of Facial Expressions*

31 Ben Ruijl (UL) *Advances in computational methods for QFT calculations*

32 Thaer Samar (RUN) *Access to and Retrievability of Content in Web Archives*